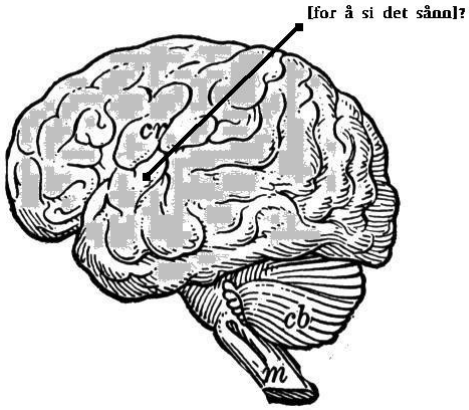
Recurrent Multi-Word Sequences and Mental Representations

A psycholinguistic perspective

Laila Yvonne Henriksen



MA-thesis

Department of Linguistics and Scandinavian studies

UNIVERSITY OF OSLO

May 2009

Preface

This MA-thesis has been funded by the *Centre for the Study of Mind in Nature* (CSMN), a Center of Excellence based at the *Department of Philosophy, Classics, History of Art and Ideas* (IFIKK) at the University of Oslo. The thesis is a contribution to the centre's research on "the interaction between the empirical study of agency and the study and articulation of the norms that govern those agents".

Some years ago I participated as an informant in the normalization of a hearing in noise test (HINT). Approximately 240 test sequences masked with noise were included, and as the testing went on, I felt that some of the sequences were read out loud and clear, while others were unintelligible because they were masked with stronger noise. Subsequently I learned that all the sequences were masked with the same level of noise. This fact made me try to recollect the sequences that I had actually perceived, and then I had a hunch: The sequences seemed familiar to me. Could it be that the sequences were frequent in use, and therefore conventional multi-word sequences, stored and processed as units?

I would like to thank my supervisor, Inger Moen, for supporting my hunch, for giving excellent advice and guidance on test methodology and for always giving me new courage after every meeting. I would also like to thank Tekstlaboratoriet for their helpfulness, and especially Joel Priestly for sharing the n-grams from NoTa-Oslo with me. Engineer Marte Myhrum at Rikshospitalet ØNH, deserves sincere thanks for helping me develop the test, and for letting me use the soundproof room at Rikshospitalet's ØNH-section. Without the informants, there would not have been an experiment. Thank you! Through the process I have been fortunate to be surrounded by my co-students and good friends, who have made the time at UiO precious. Especially Maja Sund, Eli Anne Eiesland and Ingeborg Dalby have made every single day a good day! Special thanks go to Ingeborg and Eli Anne who have read and commented my paper, and Eli Anne who also proofread the final draft. Siri Lader Bruhn has contributed with her knowledge when she read and commented on an earlier draft of the thesis. Elin Marie Strand helped me translating the data, which saved me for much work. Benedetta Frosini is part responsible for the better parts of the thesis by reading and discussing the contents, and for proofreading parts of the thesis. What would I have done without you?

My precious children Elias, Emily and Askil have been my strongest motivation for finishing this thesis, and it simply would not have been written if it weren't for the love of my life, Hans Kristian, who kept the family together and supported me all the way.

Oslo, May 2009

Laila Yvonne Henriksen

Contents

Figures

Tables

1 Preliminaries

1.1 Introduction

The most frequent five word sequence in the Norwegian oral corpus NoTa-Oslo[1] is *for å si det sånn* ('so to speak'). This sequence is one of many other recurrent and fully compositional sequences in the corpus. A fully compositional sequence is a sequence which meaning is predictable from its parts. *For å si det sånn* has been used 116 times by 50 of the 166 informants, and a search on www.google.no resulted in approximately 251.000 occurrences of the same sequence. The sequence is evidently a conventional unit, known and used by many language users. The question posed here, is whether it is stored and processed as a lexical unit in the individual language user's mind. Is there reason to assume that the fully compositional recurrent sequences in language use, observable in language corpora, are represented in language users' mental lexicons as memory units, or are these repeating patterns of language use just a matter of coincidence?

While it is fairly uncontroversial that language users store some prefabricated units, the prevailing view is that the mental lexicon of stored linguistic items is economically organized, and so, the system does not waste capacity by storing fully compositional sequences that may instead be captured by a general computational algorithm, or rule. But is this a psycholinguistically plausible assumption? The human mind is capable of remembering and automatizing all other repeated and highly specific sequences, like driving a car the same route to work every day, baking your favourite cake, or putting on make-up each morning. Why should the linguistic system behave differently by deleting specific linguistic habits from memory only because it follows general rules of syntax?

The present thesis investigates the phenomenon of recurrent sequences from a psycholinguistic perspective. The question of storage or non-storage of multi-word sequences is a new and rather unexplored field of investigation within the research on

[1] Norsk talespråkskorpus - Oslodelen, Tekstlaboratoriet, ILN, Universitetet i Oslo, available online at: <http://www.tekstlab.uio.no/nota/oslo/index.html>

formulaic language. Traditionally, most research on the subject has focused on the clear-cut cases of formulaic sequences, located at the idiomatic end on a scale from idiomatic to compositional sequences. However, lately an increased focus on the storage and processing properties of formulaic sequences has drawn attention to the borderline cases (Wray, 2002). Two basic assumptions, inherited from the generative linguistic tradition (see Section 3.5), characterize the ongoing research. The first is the assumption that there is a principled distinction between fully compositional and formulaic sequences, where the latter are assumed to be stored in the mental lexicon, while the former are assumed to be online computations with no unitary mental representations in long term memory. The second assumption is that this distinction may be captured by identifying distinct criteria that apply to sequences within each category. These assumptions leave the large group of recurrent and fully compositional multi-word sequences unexplained, because they fall into the group of online computations. Also, while generally accepted, the assumptions do in fact constitute a well-known methodological problem within the field. Different taxonomies and continua are proposed without solving the problem: What should be the criteria defining the cut-off-point between compositional and non-compositional sequences?

In the present study, the fully compositional recurrent sequences are investigated from a psycholinguistic perspective. A psycholinguistic experiment is conducted to provide evidence for holistic storage of recurrent and fully compositional multi-word sequences, and to test two competing theoretical models' ability to explain the empirical data. I argue that a usage-based model of language storage and processing is better suited than a generative dual-mechanism model to account for the pervasive use of recurrent multi-word sequences. While generative theories treat the recurrent use of fully compositional multi-word sequences as arbitrary, usage-based theories regard the language patterning as a normal and expected phenomenon.

This thesis is ultimately about language users' knowledge of grammar and the way linguists model this knowledge within mentalistic approaches. Mentalistic approaches to grammar want to represent grammars that are psychological adequate, that is, all postulated structures, principles and processes are assumed to refer to psychological entities. The mentalistic models thus give testable predictions regarding processing properties of linguistic structures.

The competing predictions deduced from an associative, single mechanism model and a dual mechanism model are the subjects of scrutiny in the present study.

1.2 Aim of the Thesis

The empirical study conducted in this thesis is intended as a contribution to the research field on Formulaic Sequences. To overcome a recognized problem of classification which is rooted in basic assumptions inherited from a traditional generative view (Poulsen, 2005), I propose a change of theoretical perspective. In the traditional view, the language system is modular and highly economically organized, which entails severe restrictions on what may be stored in the mental lexicon. While theorists within the field do not necessarily adhere to the generativist approach, the principle of economy and the division of labour between grammar and lexicon still makes its influence: Formulaic sequences are treated as exceptions to the rule, while the recurrent, fully compositional multi-word sequences, which are omnipresent in language use, represent a phenomenon that one is aware of, but which cannot be explained within the traditional theoretical framework.

While the research field in general apprehend the formulaic sequence as a storing and processing unit (see section 2.2), most research aimed at categorizing formulaic sequences and distinguishing these from non-formulaic sequences approaches their material with predefined criteria traditionally identified with formulaicity. The recurrent and fully compositional multi-word sequences, evident in language use, do not possess these properties; however, this fact does not prevent the sequences from being storing- and processing units. In the present thesis, I argue that a psycholinguistic approach is better suited to establish whether recurrent and fully compositional multi-word sequences are stored and processed as units. The thesis has two main goals:

> **The first goal is to extend the category of formulaic sequences to include conventional, recurrent, but fully compositional multi-word sequences.**
>
> **The second goal is to evaluate two competing models' ability to explain the results of the present psycholinguistic experiment.**

I will test whether recurrent multi-word sequences extracted from a Norwegian oral corpus (see Section 4.2) are likely to be storage and processing units. If this assumption is supported by the psycholinguistic experiment, it will challenge the prevalent assumption that fully compositional recurrent sequences are online computations with no representation as entrenched activation patterns in long term memory. It will also challenge the criteria traditionally used to distinguish formulaic sequences from fully compositional sequences (see Chapter 2). The aspect of compositionality is assumed to play a defining role, which then may prove to be an insufficient or inadequate method for classifying the sequences.

I propose that usage-based theory, and an exemplar model of language storage and processing are more psycholinguistically valid than models based on the assumption that the regularities of language are best kept by a separate rule based system, embracing the economy principle (see Section 1.5). While the latter theories postulate a restricted lexicon which stores only the minimal or, in some sense, irregular units (including idiomatic multi-word expressions), the former theory assumes that the recurrent sequences evident in corpora reflect both patterns of usage, and usage patterns in language users' minds, i.e. that language users store conventional multi-word sequences, because conventional multi-word sequences enhance both production and comprehension.

The more specific formulation of the thesis goals are formulated below in form of a hypothesis and research questions.

1.3 Hypothesis and Research Questions

Langacker (1987:36) states that language patterning, which is pervasive in all language use, "[is] (…) quite expected, pose no special descriptive problems, and in fact constitute a central and explicitly recognized kind of data to be accounted for". Langackers' Cognitive Grammar is a usage-based theory in which patterns of language use are viewed as reflecting the language users' grammars (see Chapter 3). Thus, independent of form and size, frequently used language structures should reflect high grade of conventionality. The conventional units in language are typically recognized and known by the language users,

which indicate that these units must have memory storage. The working hypothesis for this thesis is as follows:

Frequency of use affects the mental representation of fully compositional multi-word sequences

This entails that the recurrent use of any sequence of words, independent of the presence or absence of other properties than recurrence, leads to an imprint in language users' mental grammars. The degree of entrenchment depends on frequency, but other properties of the sequence will contribute as well (see Chapter 3). Following this assumption, the frequent sequences extracted from language corpora are alleged to constitute memory units in the mental grammars of language users that are representative of the corpus in question. In order to examine the relation between recurrent sequences and mental representation, the following research question is relevant:

1) **Do recurrent multi-word sequences have mental representation as entrenched activation patterns in language users' minds?**

According to traditional approaches within the field of formulaicity or phraseology, multi-word sequences will belong to either one of two principled different classes: Either the sequence is fully compositional, or the sequence is a lexical unit. From a psycholinguistic perspective, the sequences' status as lexical or compositional may be captured in terms of processing effort, as lexical units are generally assumed to enhance processing, while comparably, the fully compositional, i.e. new sequences demand greater processing efforts. If there is a divergence in processing efforts for the recurrent multi-word sequences versus the infrequent and supposedly non-lexical sequences, this will indicate storage for the recurrent multi-word sequences. The results from the psycholinguistic experiment make up the empirical data which theories of language storage and processing must be able to predict and explain. This leads to the next question:

2) **Is the data from the present psycholinguistic experiment compatible with a dual-mechanism model of language storage and processing, or can a usage-based, single mechanism model better capture the properties of recurrent, but fully compositional sequences?**

The motivation for choosing a usage-based approach to formulaicity is its natural inclusion of language patterning. While general research on formulaicity tends to treat the recurrent but fully compositional sequences as arbitrary products of rule based computations, these sequences are, according to usage-based theories, natural consequences of everyday language use. In usage-based models, frequency is one (amongst others) decisive factor which leads to cognitive routines. Frequent use strengthens the mental representations of these routines, enhancing their re-selection. An important functional aspect of this process is that language routines are resources for both speakers and hearers - conventional multi-word sequences ease communication by evoking commonly known linguistic patterns. The fact that usage-based language theory treats the recurrent multi-word sequences in language use as a natural and expected feature of language use and language competence makes the approach suitable for the present investigation.

1.4 The Psycholinguistic Experiment

The psycholinguistic experiment is designed to investigate whether recurrent and fully compositional multi-word sequences are easier to perceive than infrequent and presumably non-formulaic multi-word sequences. A method to test this is to compare two sets of linguistic material, one consisting of frequent multi-word constructions from an oral corpus, and the other consisting of constructed low frequency multi-word strings. By adding noise to the target sequences the speech signals are distorted, and a sequence restoration task is used to investigate if the two target groups are perceived and reproduced differently. If frequency effects are found for fully compositional multi-word sequences, this may indicate that the sequences have achieved a strong mental representation as a consequence of being frequently used.

1.4.1 Material and Design

The material used for the psycholinguistic test is the hundred most frequent four- and five-word sequences in NoTa-Oslo, a Norwegian oral corpus of interviews and conversation by 166 informants born and raised in, or in the suburbs of Oslo. Out of one hundred sequences, 30 representative constructions were chosen as targets in the test, alongside the same number

of infrequent constructions of the same size. The final test is a set of 30 frequent constructions, 30 infrequent constructions, and 62 dummy sentences to reduce priming, and practice and fatigue effects. The total set of 122 sentences was taped and masked with white noise (see Section 4.3.4). The test subjects are 30 adult native speakers with an east Norwegian dialect.

1.4.2 Data and Analysis

The taped responses of the psycholinguistic test make up the data in this study, and are quantitatively analyzed through simple statistical calculation, and qualitatively analyzed in relation to properties associated with formulaic sequences. The goals are to see if there are significant differences in the participants' reproduction between the two frequency groups, and to investigate the properties of the responses. The results are discussed in light of the two competing theories' predictions (see Chapter 5) and in relation to the thesis' hypothesis and research questions.

1.5 Terms and Definitions

In this thesis I use several different terms denoting multi-word sequences: "formulaic sequences", "recurrent sequences", "conventional expressions", and "collocations", all of which are expressions of somewhat different aspects of the phenomena in question, or tied to different theories. The least specific term is in my opinion "**multi-word sequence"**, and I use this term throughout the thesis meaning only a sequence of words, without reference to formulaicity. The term "**formulaic sequence"** (Wray, 2002) is the currently most used and accepted term denoting multi-word sequences that are supposedly storage and processing units in the language users' minds. "**Recurrent sequence"** (Schmitt et al., 2004) is a term used for the recurring sequences that are derived from a corpus. This term does not concern the notion of psycholinguistic reality. Langacker coined the term "**conventional expression"** to include the class of multi-word sequences as a whole. While I share Langacker's view of multi-word sequences, I mainly use the term "formulaic sequence" in this thesis as the term is generally accepted and refers to the storage and processing aspect of lexical sequences. The term "**collocation"** is used for the probabilistic relationship between a word and other

words that it typically co-occurs with (Biber, 2000), and so the term denotes the tendency for lexical items to co-occur in a text, or in a text corpus, whether or not they form a syntactic pattern (Poulsen, 2005: 14).

A central property of the recurrent sequences extracted from NoTa-Oslo is that they are fully compositional and analyzable. Langacker uses the term compositionality as "the regularity of compositional relationships, i.e. the degree to which the value of the whole is predictable from the value of the parts" (1987: 457). So by "**compositional"** I mean that the elements within the sequence together contribute to the meaning of the whole sequence. "**Analyzability"** refers to the individual language users' opportunity and ability to be aware of the elements within the sequence and how they contribute to the meaning of the composite whole. I also refer to the recurrent sequences as "**literal"**, which means that the sequences' composite parts are used in a literal sense. Sequences that diverge from this by being either noncompositional, unanlyzable or figurative – or all three, have traditionally been considered to be **irregular** in some sense. These irregular multi-word sequences cannot be compositionally computed or analyzed, and thus need to be listed in the lexicon as lexical units. Generative theories assume that the fully compositional sequences do not enter the lexicon as complex units. While the generalizations are kept as abstract rules in the grammar, the particular instantiations can be eliminated from memory. This "**principle of economy"** has been central in generative theories, and must be seen in relation to the criteria made by generative grammar for evaluating linguistic analyses: The optimal analysis is achieved by the application of Occam's razor, that is, the simplest analysis, which uses fewer features and rules is the preferred one (Crystal, 1997). Langacker criticizes this stand, as "[t]rue simplicity is not achieved just by omitting relevant facts. Questions of economy are meaningful raised only with reference to a particular body of data", and he states that: "It would be fallacious […] to invoke the principle of economy to argue that conventional expressions should not be listed in a grammar […]" (Langacker, 1987: 41).

In the thesis I use the terms "**processing", "perception"** and "**production"**, and since these terms denote processes that are described quite differently according to theoretical stand, I will give a short demarcation regarding my use here. A definition of "processing" is given by Field (2004: 224):

> *The analysis, classification and interpretation of a stimulus. In psycholinguistics, particularly used for the cognitive operations underlying (a) the four language skills (speaking, listening, reading, writing); (b) the retrieval of lexical items; and (c) the construction of meaning representations. The term sometimes refers more narrowly to the receptive process of listening and reading"*

In the following, I generally use the term processing to mean the above mentioned (b), "the retrieval of lexical items"; however, within usage-based models, both comprehension and production are viewed as intrinsic parts of the linguistic system (Kemmer and Barlow, 2000: xi). Thus I will use the term "processing" when speaking of the general activities executed by the processing system, and the terms "perception" or "comprehension" and "production" when speaking of the specific activities subsumed by the more general term "processing".

The term "**masking**" is used with two different meanings in this thesis: In reference to the Adaptive Resonance Theory (see Section 3.4.1), "masking" means that the largest perceptual unit attracts attention at the expense of the smaller units within the larger units. In reference to the psycholinguistic experiment (see Section 4) "masking" means the reduction or distortion of the (written or spoken) input signal.

At last, the notion of "**psycholinguistic reality"** calls for clarification. When I claim that the recurrent sequences extracted from NoTa-Oslo have psycholinguistic reality as storage and processing units in the language users' minds, it is important to stress that the unit is defined in processing terms as a cognitive routine, or patterns of mental and ultimately neural activation (Langacker, 1991:511,527), and not as static elements of linguistic structure. The linguistic units are thus viewed as highly dynamic patterns of activation, which allow redundant, interactive and distributed storage in long-term memory (see Chapter 3).

1.6 Thesis Outline

Having introduced my general research aims, and terms and definitions, I will conclude this chapter with an outline of the thesis. The thesis is divided into 6 chapters.

Chapter 2 describes the phenomenon of "formulaicity" and evaluates the research field's methods for classification and their basic assumptions about the nature of formulaic language and its mental representations.

Chapter 3 provides a theoretical frame for the thesis, contrasted with an opposing theoretical approach.

Chapter 4 includes the psycholinguistic experiment and its results, which will make up the empirical part of my thesis.

In Chapter 5, I present the interpretations and explanations of the results from this experiment in light of the research questions presented in this chapter, and in relation to the competing theoretical models presented in Chapter 4. The relation between corpus data and mental representation of language structures is discussed.

In Chapter 6, I will present my conclusions concerning the present empirical study compared to previous studies, and regarding the explanatory potential of a usage-based approach compared to a generative approach. Ideas for further research will be suggested.

2 The Research Field

2.1 Introduction

Recurrent multi-word sequences are ubiquitous in language use (Nattinger and DeCarrico, 1992: 66), and in recent years these patterns have become the subject of a fast growing research field. The focus within the field of formulaicity has mainly been on the identification, classification and delimitation of the phenomenon, and on the practical application of the research findings as a resource within language learning – how formulaic sequences are learned, both by first and second language learners, how formulaic sequences are preserved in patients with language loss, and how knowledge of this can be used to develop educational and therapeutic material (Poulsen, 2005: 37). Recent studies have also brought a theoretically oriented aspect into the research field by asking how formulaic sequences are represented in mind and how they are processed. Both approaches are relevant for the present study; however the focus will be on the latter perspective.

Within a psycholinguistic perspective, some research have been done to investigate the relation between formulaic sequences and processing demands, and the results support the general assumption that storage of larger constructions enhance processing. However, these studies have generally included the unambiguous cases of formulaic sequences; either idiomatic or irregular in some way (see Section 2.3.2). Only a very few studies (Schmitt et al., 2004; Vogel Sosa & McFarlane, 2002; Tremblay Tremblay, Derwing, Libben, and Westbury, 2008; Tremblay, Libben, Derwing, and Baayen, 2008) have investigated the frequently occurring, but fully compositional, sequences in relation to mental representation, and the results are in addition contradictory (see Section 2.3.3).

The research on multi-word sequences has grown into a specialized linguistic field. The researchers within this field agree that the language phenomenon in question is not a unitary category, and are even questioning whether it is possible to define the phenomenon in any meaningful way (Schmitt and Carter, 2004). This diversity, which characterizes the phenomenon, has resulted in several different terms denoting the phenomenon and in several definitions. One of the problems of defining formulaic sequences is the common expectation

that formulaic sequences are lexical items in the same way as words are, "and with the same properties as words would have if they were phrases" (ibid.:4). However, while words are listed in the grammar as "natural" units, frequently used sequences need additional qualities – besides being frequent, before they are considered memory units. This presupposes the existence of a principled division between sequences that are represented in long term memory and sequences that are fully compositional, and hence, supposedly need no representation in long term memory.

The goals for this chapter are to explore different approaches to formulaicity and to establish the formulaic sequence as a processing unit. A third goal is to present two studies which draw different conclusions about the relation between statistically derived material from corpora and the representation as memory units in language users' minds.

2.2 Formulaic Sequences

The heading of this section is "Formulaic Sequences"; however this term is just one of more than fifty other terms denoting multi-word units, as recognized by Wray (2002: 9, as shown in Figure 1). The vast variety of terms in use for multi-word sequences is expressive of their diversity and of the researchers' different perspectives.

Amalgams – automatic – chunks- clichés – co-ordinate constructions – collocations – complex lexemes – composites – conventionalized forms – F[ixed] E[xpressions] including I[diomd] – fixed expressions – formulaic language – formulaic speech – formulas/formulae – fossilized forms – frozen metaphors – frozen phrases – gambits – gestalt – holistic – holophrases – idiomatic – idioms – irregular – lexical simplex – lexical(ized) phrases – lexicalized sentence stems – listemes – multiword items/units – multiword lexical phenomena – noncompositional – noncomputational – non-productive – nonpropositional – petrifications – phrasemes – praxons – preassembled speech – precoded conventionalized routines – prefabricated routines and patterns – ready-made expressions – ready-made utterances – recurring utterances – rote – routine formulae – schemata – semipreconstructed phrases that constitute single choices – sentence builders – set phrases – stable and familiar expressions with specialized subsenses – stereotyped phrases – stereotypes stock utterances – synthetic – unanalyzed chunks of speech – unanalyzed multiword chunks – units

Figure 1 Terms used to describe aspects of formulaicity (in Wray, 2002:9)

Wray argues in favour of the term "formulaic sequence" because it indicates that the phenomenon in question is a *sequence* of internal units and that the sequence is both *custom* and a '*habit*' (ibid.). The term "recurrent sequences" (see Section 1.5), which I use in the title of this thesis is used in contrast to "formulaic sequences". The latter denotes linguistic structures that are stored and processed as units in mind, while the former denotes patterns of language use, which may or may not have psycholinguistic reality (Schmitt et al., 2004). While the range of different terms describe a smaller or larger part of or different aspects of related phenomena, they all have in common the fact that they denote sequences that behave as *units* in one or several ways.

The phenomenon *formulaicity* has been long recognized, at least since the mid-nineteenth-century, when John Hughlings Jackson described aphasic patients' ability to utter whole sequences of prayers, greetings and rhymes while not being able to construct novel utterances (referred in Wray, 2002: 7). The common assumption has been that the language user makes use of larger units, from collocations to whole sentences to reduce storage and processing loads. However, since Chomsky reinvented the mentalistic enterprise in the early 1960s and at the same time made a distinction between *competence* and *performance* – language

knowledge and language use respectively, this assumption was questioned.[2] Since then, facts of language use have largely been viewed as irrelevant for grammatical theorizing. Even though formulaicity has been investigated and described by several linguists continuously through this period and to our time, it has been treated as something that lies outside the linguistic field. Lately, the pervasive use of collocations and prefixed word strings evident in corpora of language use has evoked renewed interest. Alongside the new generation of grammatical theories based on performance, the research field of formulaicity is growing. It is now generally accepted that the mental lexicon of formulaic sequences must be quite extensive (Pawley and Syder, 1983; Jackendoff, 1995; Melcuk, 1995), and it is also believed that the formulaic sequences are processed more efficiently than creatively generated sequences (Pawley and Syder, 1983; Conklin and Schmitt, 2007).

One of the most extensive works on formulaic sequences has been done by Wray (2000; 2002; 2006 and Wray and Perkins, 2000). In her book *Formulaic Language and the Lexicon* (2002), Wray investigates the role of formulaic sequences in different linguistic fields like general linguistics, lexicography, language teaching, first and second language acquisition, corpus linguistics, and neurolinguistics, amongst others. Her chief interest is in the representation of formulaic language in the mental lexicon, and the processing of formulaic sequences. She defines the formulaic sequence as:

> *a sequence, continuous or discontinuous, of words or other elements, which is, or appears to be prefabricated: that is, stored and retrieved whole from memory at the time of use, rather than being subject to generation or analysis by the language grammar (Wray, 2002: 9).*

This definition is currently the most used and accepted. It has, however, also been criticized for being too broad in some respects, and, at the same time, too excluding in others. On the one hand it is all-inclusive by opening up for units of all sizes, including morphemes as well as discourse sized units. The definition states that the sequence does not have to be continuous, so there may be insertions into it, for example when the word *bloody* is inserted into *all heart*, the result is a construction which somewhat changes the original meaning:

[2] Avram Noam Chomsky (b. 1928), Professor of Modern Language and Linguistics at the Massachusetts Institute of Technology, known for revolutionizing the field of linguistics in the late 50-es, early 60-es with his book "Syntactic Structures" (1957).

you're all bloody heart, aren't you?, but is still counted as an instance of the formulaic sequence *all heart*. While the definition includes these variations, Read and Nation (2004) criticize the definition because it seems to exclude sequences which include the substitution or inflection of items within the formulaic sequence. An example is the idiom *I'll eat my hat*, where both subject and verb can be substituted by other items, for examples *Greenpeace activists say they'll eat their hats*, *hope you enjoyed eating your hat* and *he ate his hat and apologized*. Also grammatical transformations like, *Right, that's it! ....add me to the* ***hat eaters*** *list. If I am wrong, I will eat my hat* will be excluded by the definition (ibid: 26). They argue that these examples are excluded as formulaic sequences by Wray's definition because they involve "generation or analysis by the language grammar" (Wray, 2002: 9). This interpretation expresses a view which is found in Jespersen (1924, 1976, referred in Wray, 2004: 48): Formulaic sequences are sequences where no part can be changed or omitted without losing the formulaic aspect. This narrow interpretation will exclude all but some idioms, rhymes, prayers, quotations and proverbs (ibid: 49).

Wray (2002) argues that a form criterion is not the only, nor a necessary criterion for identifying a sequence as formulaic. To exemplify cases where the formulaicity is a property residing on the conceptual level rather than in the form, she uses the *is the Pope Catholic?*-construction. This construction is a rhetorical question, a sarcastic response to another question for which the obvious answer is "yes". The same effect can be achieved by other responses as for example, *do fleas like cats?* Or *does Dolly Parton sleep on her back?* While all the composite parts of the Pope-construction can be changed, the concept is preserved, something that challenges the strict form-based criteria "for what, precisely, is being stored, when all the words can be novel?" (Ibid: 32). Wray proposes that the original construction is stored, alongside alternating constructions, and that these serve as templates for both construction of new variants[3] and the recognition of unknown variants.

Regarding psycholinguistic reality, the definition of the formulaic sequences as "stored and retrieved whole from memory" (Wray, 2002: 9) is being criticized because it is difficult, or even impossible, to say that a given sequence is a property of every language user's language

[3]The construction of new alternating constructions for this kind of idiomatic expressions is a difficult and deliberate process, according to Wray (2002: 33), thus the already existing ones are stored as precious alternations to the template.

system (Read and Nation, 2004: 25). However, this aspect is problematic only if the purpose is to identify formulaic sequences in texts. A dynamic usage- based approach will not assume that all language users' grammars are identical, as "the question of storage or non-storage will always be a probabilistic one, based on the experience of the language user." (Bybee, 2006: 8) This aspect will be treated further in Section 3.3.3.

Despite the problems of Wray's definition commented by Read and Nation (2004), this definition is the most cited at the present time, and because it has its focus on the storage and processing aspects of formulaic sequences, it will be used as a reference point in the present thesis. The problems with this definition, as pointed out above, are tied to general classification problems within the field, which are the subject of the next section.

2.3 Defining Formulaicity

Definitions of formulaicity are in general based on the assumption that formulaic language is a delimited class of sequences with specific properties. Taxonomies and continua based on distinct features are different approaches which are used for describing and delimiting the phenomenon. The idea that formulaic sequences may be captured by criterial features originates from the Aristotelian theory of classification, which operates with categories of items that are defined by necessary and sufficient properties. The categories must capture all the items in question, and must be mutually exclusive, which are classification principles that have proven to be difficult to attain. The continuum models attempt to avoid the problems associated with taxonomies, but run into some of the same problems because they are based on the same criterial features as the taxonomies. While these different approaches capture the division between the clear-cut cases of lexical complex units and creative language use, they encounter problems regarding the borderline instances, not knowing how to ascribe these sequences membership to either one or the other class. The next section is devoted to an overlook of the traditional defining criteria for formulaicity.

2.3.1 Criterial Features

Different taxonomies have been proposed for formulaic sequences by several researchers (Nattinger and DeCarrico, 1992; Melcuk, 1998; Hudson, 1998, amongst others). However,

the categories are difficult to specify as they are neither discrete nor comprehensive (Wray, 2002: 46). The problem with taxonomies is the theoretical justification for a classification of formulaicity as a unitary phenomenon that can be exhaustively described and measured within one level of description. The phenomenon labelled "formulaic sequences" is not a unitary class of sequences. It is rather in fact sequences of words which vary according to different features or processes, and the different features and processes may also be partly overlapping. Thus, a sequence can be a member of several categories, or fail to be recognized as formulaic because the taxonomic categories are restricted to one mode of description. Instead of describing specific taxonomies, I will, in the following, describe features which are typically associated with formulaicity.

Wray (2002: 47) identifies four types of criteria used to define formulaicity: form, function, meaning and provenance. The **formal properties** of the formulaic sequences are hard to pin down as there is vast variation. A typical property associated with formulaic sequences is an irregularity in form. The sequences may not conform internally to the grammatical rules of the language, or they may have an unusual meaning (Wray, 2002: 49). An example is *I thee wed*, which displays a non-canonical word order. An example of semantic irregularity is the idiom *kick the bucket,* which cannot be understood online without knowing the special meaning attached to it. Syntactic irregularity is usually a consequence of language change. The formulaic sequences maintain a frozen syntax because they are stored and processed as units, whereas structural changes affect the combinatorial possibilities for simple forms. The typical formulaic sequences are also less variable and more continuous than compositional sequences. They allow fewer morphosyntactica adjustments and also fewer slots within the sequence where words or phrases can be inserted. However, according to Wray (2002: 50), the totally unchangeable formulaic sequences, where one cannot replace or delete anything within the sequence without changing or ruining the original meaning, are exceptions. It is therefore difficult to use these criteria other than to identify idiomatic expressions, which are only a small subset of the class of formulaic sequences. Other criteria described below capture a broader class of formulaic sequences.

Function is a much applied criterion. Coulmas (1981: 2f) describes the formulaic sequences as "expressions whose occurrence is tied to more or less standardized communication situations." This criterion may be useful to identify the functions connected to formulaic sequences. Like words, formulaic sequences typically represent specific conceptual

meanings. The derivation of a specific formulaic sequence's meaning is thus fairly unproblematic. However, the opposite direction of derivation is more difficult because a specific function may be achieved by using one of several sequences (Wray, 2002). There is thus not a one-to-one relation between functions and linguistic units. Many sequences that do not qualify as formulaic by other measures may be captured by this criterion, so the criterion reaches across a larger area of sequences. Still, this criterion excludes sequences which are recurrent, but which have not specific pragmatic meanings attached to them.

Traditionally the most applied criterion to define formulaicity, besides irregular syntax, has been **meaning**. The sequences that are not fully analytical must be stored in the lexicon, because they have meanings that cannot be reached through analysis. Thus, successful processing of these sequences demands that they be accessible as units. This criterion, however, is only useful to identify idiomatic expressions, excluding strings with literal, referential meaning, which may still be formulaic because they have a specific pragmatic meaning or because of other properties.

Formulaic sequences are often defined from the way they come about. Wray (2002) uses the term "**provenance"** for this process, and recognizes three different paths: First, a sequence may be adopted by the language user as a formula, and it may or may not be broken down into its composite parts by the user. Wray gives the example *rice crispies*. Many language users have learned this sequence as a unit, and for some it has obviously never been analysed, as they expressed surprise upon learning that the cereal was made of rice (Wray, 2002: 3). Second, the sequence may start off as a creatively generated sequence that becomes formulaic as a consequence of adopting a different meaning, either semantic or pragmatic, and therefore is stored as a unit. Third, the sequence may become formulaic as a consequence of repeated use: it fuses and becomes automatized. Wray assumes that the first two processes are primary, and criticizes Langacker (1987), Bybee (1998) and Lamb (1998) for a one-sided focus on fusion (Wray, 2002: 273f). A discussion of these opposing views is given in Section 5.1.1.

While taxonomies are categorical, formulaic sequences have also been described as lying along a continuum from less to more formulaic. Pawley and Syder (1983) propose a "novelty scale" that focuses on variability from entirely novel to entirely memorized sequences. This continuum opens up for intermediate cases consisting of partly new collocations of lexical

items and partly memorized lexical and structural material. Another continuum in the same vein as Pawley and Syder's has been proposed by Howarth (1998). This continuum ranges from fully compositional at one end to the most fixed idioms at the other. Howarth argue that the cut-off-point between formulaic sequences and non-formulaic sequences is between the fully compositional sequences and sequences that contains one or more items which are restricted by semantic or syntactic features (Poulsen, 2005: 78). Combinations like *writing a letter*, etc. should not be counted as formulaic because they "(…) pose no problems for learners. Although recurrent and familiar, they are composed according to standard rules of syntax and semantics" (Howarth, 1996: 181, referred in Poulsen, 2005: 83).

While Pawley and Syder's, and Howarth's continuum models focus on the formal properties of the linguistic material, Givón's (1989) model focuses on the processing of sequences, and proposes an "automaticity continuum". This continuum ranges from the most conscious processing that demands high grade of attention, to the most automatic processing that includes more predictable tasks. A problem with the continuum models according to Wray (2002: 63) is that from the hearers' view it is difficult to see how a sequence may hold an interim position; either the sequence is processed formulaically or it is not. Wray claims that formulaic sequences are clear-cut cases and solves the problem of identifying the formulaic sequences by focusing on the processing aspect. Sequences that are processed as units are formulaic. She argues for two separate processing systems, one for holistic processing and the other for synthetic processing. The question of formulaicity is thus tied to the individual's memory system, which either has, or has not internalized the sequence as a unit. This view is also held by Sinclair (1991), who divides the processing system into two separate and incompatible principles of computation. The "idiom principle" selects sequences of two or more words together from the lexicon, which is the primary choice in processing. However, a lexical choice which is unexpected in its environment causes a switch to the "open choice principle".

One of the properties that Wray (2002) and Givón (1989) emphasize as an attribute of the formulaic sequence is that it is a storage and processing unit. Some work has been done to investigate whether the storage of formulaic sequences contributes to a reduction in processing-load compared to non-formulaic sequences, and this assumption has indeed been supported by several studies (Pawley and Syder, 1983; Wray, 2002; Conklin and Schmitt, 2004; Schmitt and Underwood, 2004; Underwood et al., 2004, Trembley et al. 2008a). This

property of the formulaic sequence is held as a basic assumption in my thesis, and is used to justify the research design in 4.3. The next section describes selected research within the field of processing of formulaic sequences.

2.3.2 A Psycholinguistic Perspective

Despite the tendency for treating formulaic sequences as arbitrary deviations from regular compositional language, the processing benefits of storing larger constructions are intuitive. Instead of using a lot of processing effort when constructing or decomposing novel sequences of words, we benefit from following established language routines. So instead of asking: *would you like to go to the canteen and drink a cup of coffee with me in five minutes?* the shorter, conventional sequence *coffee in five?* with a nod in the direction of the canteen is used, again and again, alongside other variants that are more or less conventionalized. Several studies (reviewed below) have demonstrated that formulaic sequences have processing advantages over non-formulaic sequences, and so, there is now generally a consensus that formulaic sequences have a significant function in minimizing processing efforts for both speaker and listener: Following conventional language routines enhances communication.[4]

The question about the way in which formulaic sequences are processed compared to non-formulaic sequences is central to the present thesis, and the remainder of this section is therefore devoted to a short review of two studies that use different methodologies to explore the processing properties of formulaic sequences. Both conclude that the formulaic sequences are, or seem to be processed as units, and that the formulaic sequences reduce the processing load compared to non-formulaic sequences.

Underwood et al. (2004) tested processing of formulaic sequences in a reading task, working from the hypothesis that the final word in an idiom requires less attention, as the whole phrase has been recognized from the first few words. Eye tracking methodology was used to determine whether the final words of the idiomatic formulaic sequences were fixated upon for a shorter time than the same words in non-formulaic sequences. They chose to use

[4] The notion of conventionality will be elaborated in Chapter 3.

unambiguous cases as targets in the test. The results showed evidence for a processing advantage for the formulaic sequences relative to the non-formulaic sequences. The results do not however say if the sequence was processed as a unit, or if it was processed word by word until the sequence was recognized as an instance of a remembered sequence, causing the reader to omit the last part, or parts of the sequence. This aspect will not be treated further here, as it is not very relevant for the present thesis.[5]

Another reading task conducted by Conklin and Schmitt (2007) investigated whether formulaic sequences were read more quickly than equivalent non-formulaic sequences. Their assumption was that formulaic sequences are processed as chunks rather than word by word. To assure that the sequences indeed were formulaic, the target sentences in this experiment were mainly idioms, representing idiomatic meanings that cannot be derived from the sequences' composite parts. The formulaic sequences were analysed according to frequency based on the British National Corpus, and candidates with relatively low frequency were excluded, resulting in a list of the twenty most frequent and best-known formulaic sequences as targets. The same number of control phrases were constructed by rearranging the words in the formulaic sequences, for example, *hit the nail on the head → hit his head on the nail,* which cannot be interpreted with the idiomatic nor the literal meaning of the formulaic sequence. The twenty formulaic sequences were inserted in texts that supported the idiomatic interpretation. The same sequences were also inserted in texts that supported the literal meaning. 19 native English speakers and 20 non-native English speakers from different L1s were chosen as participants.

The results for the native English speakers showed that the reading times for formulaic sequences in a context that supported the idiomatic reading was significantly shorter than for the reading times for the controls. The results also showed that the reading times for the formulaic sequences in a context that supported the literal meaning were also significantly shorter than for the control phrases. However, there was no significant difference between reading times for the formulaic sequences in different contexts supporting either the idiomatic or the literal interpretation. The same results were found for non-native speakers.

[5] In the thesis I use the term processing unit; however, the possibility that strong entrenchment leads to faster retrieval rather than unitary retrieval is just as feasible. See MacWhinney (2000) for a discussion of this aspect of holistic processing.

They concluded that their results support the assumption that formulaic sequences are involved in more efficient language processing, whether they are used idiomatically or literally. This last fact challenges Wray's presumption that the literal reading of an ambiguous idiomatic/non-idiomatic sequence is always processed analytically, while the idiomatic reading is always processed holistically (cf. Section 2.3.1).

These studies, like the majority of studies investigating possible processing benefits for holistically stored sequences, have mainly focused on, and made use of, the unambiguous cases of formulaic sequences. Lately, some studies have directed their attention to the borderline cases. These studies take as their starting point the recurrent multi-word sequences of different types. The assumption that recurrent sequences extracted from text are "stored as holistic formulaic sequences in the mind" (Schmitt et al. 2004), as these sequences can be seen as reflecting the underlying mental patterns of the language users who have produced them has not, that I am aware of, been extensively explored. Two studies that explore this assumption are presented in the next section, which questions whether frequency is a defining criterion for formulaicity.

2.3.3 Is Frequency a Significant Criterion?

In her book about formulaic sequences, Wray (2002) emphasizes the advantage of computer searches in corpora to recognize patterns of language use. Unlike introspection, a computer search misses nothing, so even patterns of ordinary language use which often are not apprehended by introspective methods, are captured (Wray, 2002: 26). Despite the advantages of this method, she stresses that frequency-based analysis is far from sufficient when it comes to identifying formulaic sequences. This is an undisputable fact, as infrequent sequences may still be formulaic. And what's more, no matter how large the corpus, it still may fail to contain well known formulaic sequences. An example is the well known idiom *kick the bucket,* which in fact was not represented once in a written English corpus of 18-million words (Moon, 1998). While corpus analysis may not alone be suited for identifying formulaic sequences, the relevant question here is whether the recurrent sequences extracted from corpora are storage and processing units as a consequence of being frequently used. Wray states that even if it can be assumed that strings that are frequently used are likely to be stored as units, "it is not possible to assert that all frequent strings are prefabricated" (Wray,

2002: 31). She claims that a list of complementary criteria is needed to distinguish between frequent sequences that are formulaic and those that are not. Wray's assertion presupposes that a principled distinction can be made between sequences that are processed by the application of well defined rules, and sequences that are stored and processed holistically (Wray, 2002: 14; 278f), and that specific properties of the sequences, besides recurrence, are necessary to give the sequences status as storage and processing units.

Two studies that examine the relation between recurrence in text and mental representation are Schmitt et al. (2004) and Sosa and MacFarlane (2002). The studies will be briefly summarized below. These studies came to conflicting conclusions about both the relation between recurrent sequences and holistic storage, and about the role of frequency. The studies serve as reference points for the empirical study in Chapter 4, which is comparable to these studies.

"Are corpus-derived recurrent clusters psycholinguistically valid?"

The study by Schmitt et al. (2004) was motivated by the common sense assumption that recurrent patterns of language use, evident in language corpora, should reflect holistic storage in language users' minds. The study was designed to investigate if corpus-derived recurrent clusters are psycholinguistically valid in terms of holistic storage. Schmitt et al. created a list of corpus-derived clusters by extracting three and four-word recurrent clusters from *Longman Grammars of Spoken and Written English* (Biber et al., 1999, referred in Schmitt et al., 2004) and *Lexical Phrases and Language Teaching* (Nattinger and DeCarrico, 1992). They took words from Hyland's (2002, referred in Schmitt et al., 2004) list which are used to express doubt and certainty, and words that are used as discourse markers. The words were then submitted to a corpus analysis to see if they formed a head in a formulaic sequence. The list of recurrent clusters was checked for frequency in three corpora: a corpus of written English, The British National Corpus (BNC), an oral English corpus, the CANCODE corpus, and an academic spoken English corpus, MICASE. From the list they then selected 25 target clusters that were balanced according to length, frequency, transparency of meaning, type of cluster, and according to the researchers' intuition about which clusters seemed likely to be stored as units.

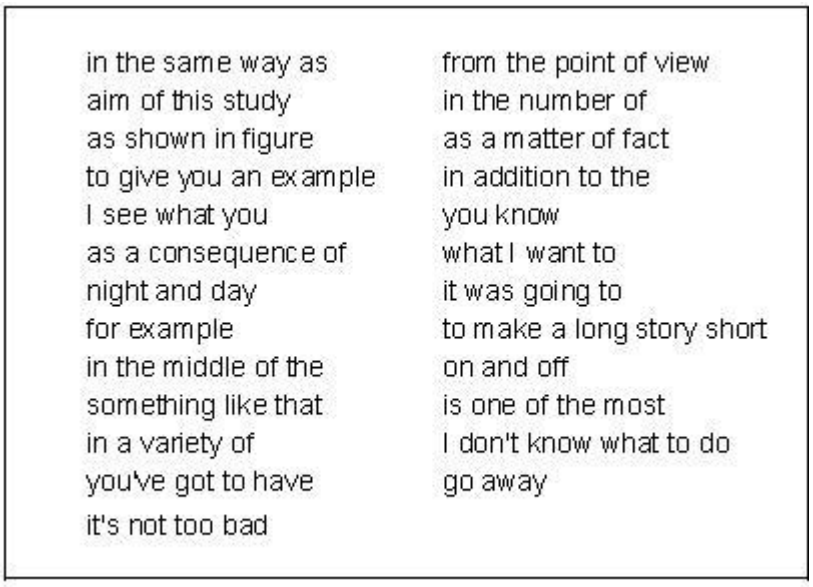


Figure 2 List of recurrent sequences used in Schmitt et al.'s (2004) study

The test methodology was chosen from the field of second language measurement. They inserted the recurrent clusters in a dictation test with the assumption that if the dictation stretches, or "bursts", were long enough to overload the working memory, the recurrent clusters had to be reconstructed from memory rather than just be repeated from rote memory. Schmitt et al. predicted that the recurrent clusters represented as storage units in mind should be produced both holistically and in a fluent manner as part of the participants' responses. They chose to construct a coherent text as the dictation bursts' context; a story about a hitchhiker. The complete story was recorded with a half minute pause between each dictation burst so the participants were allowed to complete their tasks.

The participants consisted of two groups; one group of 34 native speakers, and the other group of 45 non-native speakers with different L1s. The dictation task was sufficiently challenging for the non-native participants; however it was too easy for the native participants who repeated the complete text almost without mistakes. The researchers therefore inserted an extra, basic addition task to make the pressure on the working memory harder. The final version, after the piloting process, had 39 bursts in total, with 25 bursts containing target clusters. One of the dictation bursts are given as an example below:

The hitchhiker kept talking. Did you know there has been a sharp increase ***in the number of*** *teenagers driving drunk?* ***36+45*** (the target cluster and the additional task in bold).

The results of the experiment were classified into *mean performance*, *produced correctly, produced partially incorrect,* and *not produced*, and the distribution was interpreted such that the recurrent clusters fell on a cline of probability to whether they were stored or not in the mind of proficient speakers. They explain this with the fact that every person has their own idiolect: most peoples' lexicons probably contain the formulaic sequences that are judged to be more conventional, whereas the less conventional recurrent clusters are not represented in as many lexicons.

Schmitt et al. also investigated if there were attributes of the recurrent clusters that might affect their probability of being storage units. Besides frequency, the length of the clusters and the transparency of their meaning/function were explored. The researchers found no correlation between frequency of occurrence and performance on the dictation task, and they conclude that frequency is not closely related to whether a cluster is a storage unit or not. The same result was found for the length of the clusters; though they found that semantic and functional transparency did have a somewhat stronger relation with the performance score than frequency or length.

They concluded that corpus data are useful to identify the recurrent clusters in a language; however they hold that the results of their investigations suggests that one cannot posit a claim that the recurrent clusters are also storage units in mind just because they are recurrent.

This conclusion is counter to my hypothesis: they find no correlation between frequency and performance score for their participants, which is exactly the relation I expect to find in my experiment in Chapter 4. In Chapter 5, I argue that the methodology in Schmitt et al. (2004) is problematic in several ways, which in my opinion leads to contradictory results compared to the study by Vogel Sosa and MacFarlane (2002), which is presented below.

"Evidence for frequency-based constituents in the mental lexicon: collocations involving the word of"

Vogel Sosa and MacFarlane investigated, like Schmitt et al., whether larger units like collocations and phrases may be stored and accessed holistically. They suggested that the mechanism that determines the storage as holistic mental representations is the collocational frequency, that is, the frequency with which items occur together in natural, connected speech (Vogel Sosa and MacFarlane, 2002: 227). To test this hypothesis they selected 24 test

utterances containing a collocation involving the word *of*. The test utterances were divided into four levels of collocational frequency with six collocations in each group, listed in Table 1 below:

Table 1 Four levels of collocational frequency (in Vogel Sosa and MacFarlane, 2002: 231)

Four levels of collocational frequency			
Group 1 1→99	Group 2 100→299	Group 3 300→799	Group 4 800→Above
Mean = 24.33 Range [1, 58]	Mean = 224.5 Range [175, 281]	Mean = 391.17 Range [308, 479]	Mean = 1841 Range [889, 3592]
Sense of	Care of	Couple of	Kind of
Piece of	Because of	Part of	Lot of
Sums of	Kinds of	Most of	One of
Each of	Bit of	All of	Out of
Example of	Any of	Think of	Sort of
Colleague of	Much of	Type of	Some of

45 participants, all fluent in English were tested individually in a word-monitor task for the word *of*. The hypothesis was that the reaction time to utterances containing collocations of high frequency should be slower, as the compositionality within the sequence is reduced as a consequence of frequent use, both phonologically and semantically, and therefore the element *of* should be more difficult to spot.

The results show that the mean reaction times to the high-frequent collocations in Group 4 (see Table 1) are significantly higher than the reaction times to the less frequent collocations. They conclude that the results indicate storage of the two-word collocations in the high-frequency group, and that the significance can be attributed to frequency effects.

In line with Vogel Sosa and MacFarlane's approach to collocations, Poulsen (2005) sees conventional collocations as language routines which help the speaker to guide the hearer by calling on well known cognitive routines (Poulsen, 2005: 285). In her doctoral dissertation she gives a detailed treatment of semantic and functional properties of conventional collocations. Vogel Sosa and McFarlane, and Poulsen, also supported by Tremblay, Derwing, Libben, and Westbury (2008) and Tremblay, Libben, Derwing, and Baayen (2008), argue that a usage-based theoretical framework is better suited to describe and explain language patterning than the traditional approach. Usage-based theory, which is the

theoretical framework for the present investigation, and a competing formal theory are outlined in the next chapter about theoretical reflections on storage and processing in relation to formulaicity.

3 Language Storage and Processing

3.1 Introduction

In this chapter, I present the theoretical framework for the present thesis. A usage-based theory of the linguistic system and a usage-based exemplar model of speech perception serve as the theoretical backdrop to describe and explain the recurrent use of multi-word sequences. A usage-based approach to multi-word sequences focuses on the inseparability of lexicon and grammar. Lexical units are units of any size that represent the language users' knowledge of linguistic convention, and, as Langacker (1987: 36) states it: "much of this knowledge resides in his mastery of conventional expressions".[6] According to usage-based theory, there is no principled way to divide fully compositional sequences from irregular sequences, because the syntax-lexicon interface is constituted by a continuum of symbolic structures (ibid.), where the fixed expressions of the language, regardless of generality or size, are represented at the lexical end while the creative or free combinations are represented at the grammatical end. Within usage-based theories, frequent use predicts a strong mental representation of language structures no matter what size they are.

I contrast the usage-based approach with a generative dual-mechanism model (see 3.5.1). These theories agree on some matters; both view the lexicon as a distributed and associative memory system. Even so, they disagree about 1) how the linguistic system is structured, 2) what belongs in the mental lexicon, and 3) how frequent, regular multi-word sequences are computed. Contrary to the usage-based view, the generative dual-mechanism model assumes a distinction between lexicon and syntax. The lexicon is held to be economically organized (see Section 1.5), consisting of the minimal units in language and complex units which cannot be comprehended or generated by use of general syntactic rules. Within this view, the frequent and fully compositional sequences are assumed to be online products of rule-based

[6] Langacker uses the term 'conventional expressions' to cover all types of multi-word sequences, including the fully compositional sequences.

computations, which obstruct the sequences from being listed in the lexicon. Lexical units, but not the fully compositional units, are assumed to be affected by frequency of use.

The chapter circles around tree major themes: Firstly, how the contrasting theories view storage of multi-word sequences. This includes their assumptions about the extent of lexicalized multi-word sequences and the principles governing the lexicalization of linguistic structure. Secondly, how the competing theories model the processing of multi-word sequences, and thirdly, how these different assumptions give competing predictions regarding perception of the regular recurrent sequences, extracted from the Norwegian oral corpus No-Ta-Oslo (see Section 4.2).

3.2 Usage-Based Theory

Usage-based theories, especially associated with researchers like Langacker (1987; 1991), Bybee (1985; 2001; 2007); Croft (2001) and Tomasello (2003) are all committed to the usage-based thesis, in which "Language structure emerges from language use" (Tomasello, 2003: 5). In this view, the patterns of language use give rise to the internal language structures, that is, the mental grammar.[7] The grammar is viewed as a highly dynamic system, both constituted and changed by use in a kind of feedback loop (Barlow and Kemmer, 2000: ix), illustrated in Figure 3 below:

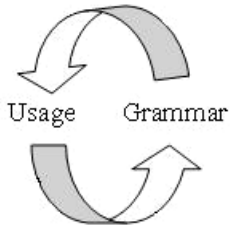


Figure 3 The intimate relation between usage and grammar

The usage-based linguistic system is massive and highly redundant, which means that it is based on the instantiations of language use without assuming a reductive device. The system

[7] Usage-based models are not restricted to a cognitively-oriented view of the linguistic system (Barlow and Kemmer, 2000), however I focus on this branch which is more relevant for my thesis.

stores all form-meaning pairs, even if an analysis results in multiple representations of the same units. To give an example, the high-frequent sequence *for å si det sånn* ('so to speak') is assumed to be stored as a unit, and so are the sequence's component parts. The words *sånn, det, si, å* and *for* (*there, it, say, to*, and *for*) are represented in the lexicon as units within other sequences, and as independent units which can be combined into new sequences (Bybee, 1998: 435; 2001: 161). This view is a maximalist one, lacking the methodological principle of economy (cf. Section 1.5) associated with generative theory. While the generalisations in language are captured by, and reduced to, mental rules within generative theories, these rules have no usage-based theoretical counterpart. Instead, the generalizations are viewed as organizational patterns, or schemas, which emerge from the instances of use. This is a bottom-up process where the general properties arise out of the specific, thus the schemas are assumed to have no existence independent of their instantiations (Bybee, 2001; Langacker, 1987; 1991).

Usage-based theory does not assume a division between grammar and lexicon: "Lexicon, morphology, and syntax form a continuum of symbolic structures, which differ along various parameters but can be divided into separate components only arbitrarily" (Langacker, 1987: 3). It still makes sense to speak of lexical units: Lexical units are linguistic structures that are mastered to the degree that the language user can comprehend or retrieve them without having to focus on their individual parts or their arrangement (ibid.: 57). However, Langacker (1987) states that there is no sharp division between units and non-units. With this in mind, the mental lexicon is depicted in the next section.

3.3 The Mental Lexicon

Langacker (1991: 549) defines the mental lexicon as "the set of fixed expressions in a language". Traditionally, these expressions have mainly been restricted to words (Carrol, 2004; Field, 2004) – or "*lexemes",* which is a term denoting the abstract units underlying the grammatical variants, like *walk* in *walks, walking, walked*. The lexemes are the minimal distinctive units in the language system, and therefore are the idiomatic sequences, which are defined as semantic un-analysable complex units, also counted as lexical units (Crystal, 1997). According to this view, everything that can be derived by the application of a generalized rule can be excluded from the lexicon (Evans and Green, 2006). Hence the role

of lexicon, in Chomsky's (1972: 39) words is to: "specify just those properties that are idiosyncratic, that are not determined by linguistic rule". Much has happened within grammatical theories since then; still, the economy principle is maintained within generative oriented theories.

Usage-based models, on the contrary, see lexical units as emergent from language use.[8] This allows instances of use to be represented as exemplars in memory, or instances of use affect the exemplars they apply to by means of similarity. Several studies support this assumption and provide evidence for lasting and detailed memories for linguistic processes (Hintzman et al., 1972; Schacter and Church, 1992; Church and Schacter, 1994; Palmeri et al., 1993; Goldinger 1996, all referred to in Goldinger, 1998). Goldinger (1998; 2004) represents a usage-based exemplar view, in which "[d]etailed episodes constitute the basic substrate of the mental lexicon" (Goldinger, 1998: 251). He asserts that memories of linguistic episodes reflect their associations to other similar memory traces in addition to the unique details of the specific episodes.

An example is in order: If you hear a specific sequence, i.e. *Next stop is [place],* a sequence that most likely is repeated several times within a short time span, if you are located on a bus that is. The next time you are on a bus, you are likely to recognize the sequence. Most likely you will even come to expect the utterance, and maybe you will be surprised if it is not uttered. Then, if you hear the sequence in another setting, for example *Next stop is bed,* uttered by a friend after a late night at the pub, you will probably recognize the sequence and notice that it is uttered out of (the ordinary) context, and appreciate the ingenuity. While the sequence is the same in both contexts, the latter expands the meaning potential. A search on www.google.com resulted in approximately 481.000 occurrences of *Next stop is*. Some of the instantiations actually referred to the "public transport frame", i.e. *Next stop is NYSUT for busload of second graders.* This instance is however referring to the "public transport frame" in an indirect way, as the school bus actually had a breakdown at NYSUT – it was not supposed to make a stop at that location. Most of the instantiations, however, include a transferred meaning of location, i.e. *Next stop is the final round in Geneva; Next stop is freedom; Next stop is the Olympics.*

[8] This applies to syntactic patterns - or constructions as well; however this is less relevant for the present thesis.

In order to associate the different meanings of one specific sequence, the sequence must be represented in memory in association to previous uses. Bybee (2007: 290) states that: ”It is my view[…] that fixed “lexical” alternations occur only within storage and processing units and could not be maintained if they were indeed applying across boundaries between processing units.” If the sequences are reduced to minimal units of meaning, there is no way that polysemous networks of different meanings, associated with the same sequence, can emerge, nor that the original meaning can contribute to new uses.[9] The notion of polysemy in the lexicon is thoroughly investigated in Tyler and Evans (2003), where they present a model of principled polysemy. While this model is applied to single words in this case, the model is assumed to include all linguistic units, from morphemes to larger constructions.[10] Frequently used lexical units are typically polysemous, where the different senses form a complex category (Langacker, 1991: 4).

According to usage-based theory, the mental lexicon is not restricted to knowledge of lexemes (see “lexemes” above). Knowledge of language is represented in associative networks of activation patterns, and these networks represent phonologic and syntactic, as well as semantic/pragmatic and social aspects of thought, because, as Goldberg (1995: 5) states it: “Knowledge of language is knowledge”. Following Goldberg (1995; 2006), language is not a cognitive capacity with its own set of properties and principles. The linguistic system is viewed as not sharply divided from other cognitive systems. This is why linguistic units cannot be categorically distinguished from non-linguistic units. For example, there may be reason to ask whether gestures are linguistic in nature or not. The principles that affect and organize general knowledge also affect and organize linguistic knowledge. According to usage-based theory, the recurrent sequences in language use reflect general properties of the cognitive system. The central properties and factors assumed to affect linguistic units are entrenchment, salience, attention, automatization, frequency and conventionality, which I will go through in the next sections.

[9] “’Polysemy’ is the phenomenon where a single linguistic unit exhibits multiple distinct yet related meanings” (Evans and Green, 2006: 36).

[10] Tyler and Evans polysemous network included English prepositions.

3.3.1 Entrenchment

Within usage-based theory, all experiences are assumed to create neuro-chemical traces. This cognitive process is termed "entrenchment" (Langacker, 1987; 1991), "memory strengthening" (Bybee, 1998; 2001), or "automatization"; though the last term refers specifically to the result of repeated use. Because repeated use creates increasingly deeper traces, this aspect is central, however, there are other routes to entrenchment as well. Unique and salient units (see Salience and attention below) may be strongly entrenched even though they might be low in frequency. Nonetheless, it is plausible to assume that some kind of entrenchment threshold exists, thus some grade of repetition must take place before activation patterns are represented as units in memory.

Units may be more or less entrenched. Usually, units get highly entrenched as the result of repeated use. Every time a pattern is activated, the memory representation is strengthened, which again facilitates its reoccurrence (Langacker, 2000: 3, cf. the feedback loop, illustrated in Figure 3 above). The theory of entrenchment omits the problem of defining the formulaic sequence as an either-or-category; instead it is a more-or-less kind of mechanism: Even if a word or a sequence has not yet achieved conscious unit status, it has to be represented in the mind as a neuro-chemical trace, which at least has the potential to become an accessible unit.

Frequency of use has been mentioned as one factor leading to entrenchment, and more will be said about frequency, which is a central theme in this thesis. However, first, a short description of the notions of salience and attention is in order, since these notions are important factors in relation to the lexicalization process.

Salience and attention

Being salient is defined as "having a quality that thrusts itself into attention" (www.freedictionary.org). Like entrenchment, salience is gradable. It is determined by factors like conventionality, frequency, familiarity, or prototypicality (Giora and Fein, 1999: 243). An example is the ambiguous word *game,* which has at least two salient meanings: 'something to play' and 'large, wild animal, and its meat'. Which of these meanings is the most salient depends on language users' knowledge of the word's senses, which to a great

extent depends on the frequency of each sense of the word.[11] Most English speakers know the sense of "game" that refers to 'something to play', while the sense that refers to 'an animal living out in the wild' does not come just as easily to mind, except for hunters, that is. The reason is presumably that the word is more often encountered in the context of playing, than of shooting or eating. Salience is a property of the lexical units, and it is especially evident considering mental processes, as salient units achieve attention more easily than less salient units. This explains how low-frequent units may still be easily perceived or accessed in processing, especially if the context contributes to prime the unit in question.

I opened this section by defining salience as a quality of memory units that attracts attention. The opposite may also be claimed: If a new word or sequence achieves attention, it may be stored in memory as a salient unit (Logan, 1988). Every day we are bombarded with impressions from our surroundings. It is essential to be able to block much of the impressions, but what is even more important is the ability to focus our attention when needed. This blocking and focusing of attention is evident in different people's memories of specific events: while one person recalls certain aspects and details of an event, another person may remember quite different aspects and details of the same event. The psychological principle of attention thus plays an important role in the organization of the mental lexicon, and can to some extent explain individual differences. Nonetheless, the assumption is that language users' mental lexicons share knowledge of the world and knowledge of language. Besides being humans and, if located at the same places, experiencing much of the same, the frequency with which these experiences, including language experiences, occur, affects our mental representations of them.

3.3.2 Frequency

Within usage-based theory, frequency is viewed as a fundamental factor in shaping the language system, because frequency is assumed to correlate with entrenchment.

[11] The word "sense" is used here to mean 'meaning'.

> *If the language system is a function of language use, then it follows that the relative frequency with which particular word or other kinds of constructions are encountered by the speaker will affect the nature of the language system (Evans and Green, 2006: 114).*

It is the frequency of use of a given word or sequence, for the individual language user, which leads to entrenchment of that word or sequence. Thus, a specific unit's frequency in a corpus and its mental representation is a probabilistic relation, depending on each language user's personal experience (Bybee, 2006: 8). Frequency data in language corpora are more closely connected to the notion of conventionality, which is a concept I will elaborate in Section 3.3.3.

Two main types of frequency are token frequency and type frequency. **Token frequency** strengthens the memory representations for instantiations of language use. For example, the repeated use of the sequence *for å si det sånn* ('so to speak') will make the sequence be entrenched more deeply, and the stronger the representation, the more probable is it that the unit will be activated in future speech events. **Type frequency**, on the other hand, is the generic result of the mutual activation of several associated instantiations, which leads to the apparently abstract schemas[12]. For instance, the words *hugged, kissed and loved* can be captured by the past tense schema [VERB-ed] (Evans and Green, 2006: 119) – or the whole sequence *loved, hugged, and kissed* may be captured by the schema [VERB-ed, VERB-ed, and VERB-ed] if it is sufficiently used and extended to other sequences fitting the same schema.

Frequency effects on linguistic elements are attested in several studies by Bybee (1998; 2006). She recognizes three effects of token frequency depending on the grade of frequency. According to Bybee, frequent constructions may become 1) conserved, 2) autonomous and/or 3) reduced. **Conservation** is the case where high frequency constructions resist morphological and syntactic changes because their morphosyntactic structures are deeply entrenched. An example is the high frequency irregular past tense verbs in Norwegian, for instance: *var* ('was')*; fikk* ('got')*; så* ('saw'), which resist regularization. The low-frequency

[12] This description of type frequency reflects an exemplar view associated with linguists like Goldinger (1998, 2003) and Bybee (2006).The question of abstraction from instantiations to abstract schemas will not be explored further, since this question is not essential for the present thesis. The choice of an exemplar-based presentation here is motivated by the personal preference of an exemplar model of perception (see Section 3.4.1).

past tense irregular verbs, however, tend to regularize, for example: *sverget/**svor*** (‘swore’) and *gravde/**grov*** (‘dug’), where the irregular forms (in bold) are about to be suppressed by the more frequent regularized forms *sverget* and *gravde*. Still, the irregular forms are conserved to some degree when they are part of formulaic sequences, as in *bannet og svor* (‘cursed and swore’) and *spurte og grov* (‘asked and dug’). These formulaic sequences occurred 208 and 1130 times respectively in a search on www.google.com, while the same constructions including the regularized forms of the verbs, *bannet og sverget* and *spurte og gravde* occurred 122 and 527 times, respectively.

Elements within morphological complex constructions or sequences of words that are extremely frequent, may move away from their original – or etymologically related meanings and become **autonomous** from the original base forms. The reason is that the complex forms are more frequent compared to the original base forms. Bybee (2006: 6) gives the example *be going to,* where the verb *go,* in this construction, does not (usually) denote the action of walking, but a transferred meaning of intention (Bybee, 2006: 15), for example *I am going to twitter this Idol madness.*

The most relevant frequency effect for the present study is **phonetic reduction**. Bybee (2006) explains this effect as the result of repeating sequences of neuro-motor routines. The execution of sequences becomes more fluent as the sequences are repeated. Because the reduced sequence is also affecting the memory representation (cf. the feedback loop, Figure 3), every new repetition leads to further reduction. The phonetic reduction and the resulting reduced compositionality is evidence that the sequences are represented as units (also supported by the study by Vogel Sosa and MacFarlane (2002), outlined in Section 2.3.3). An example is the extremely high-frequent sequence *det er ikke* (‘it is not’), which is often reduced to /dæke/ in natural speech. The reduced form is even attested in orthographic use. These are examples from www.google.com: *Dække mulig!* (It’s not possible!’)*; Dække sant?* (It’s not true?)*; Dække lov med noe sniking* (it’s not allowed to sneak [in]’).

In a usage-based lexicon (cf. Section 3.1), "memory for language consists of a large store of units of varying sizes (from word to phrase or even clause) with varying degrees of strength, productivity and connection with other units" (Bybee, 1998: 422). The units and their connections form an associative network, illustrated in Figure 4 below:

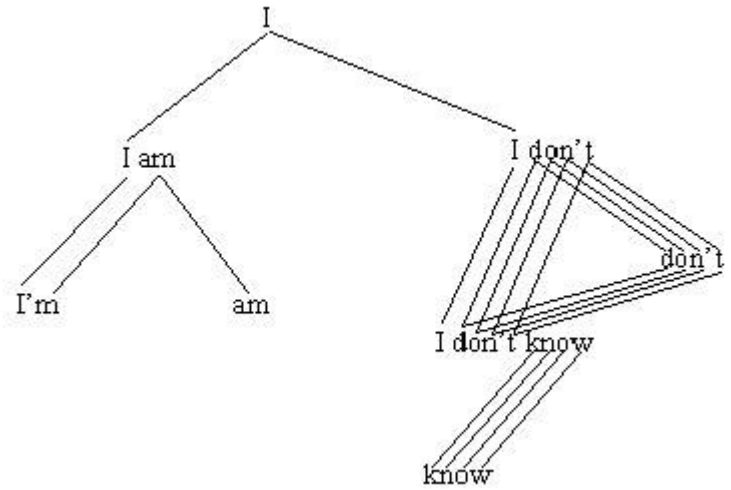


Figure 4 An associative network containing redundant storage units (in Bybee, 1998: 426)

Automatization

The proverb "practice makes perfect" is expressive for the process of automatization. Every process being repeated will eventually become automatized, including language production and retrieval. Logan (1988: 493) states that automatization reflects a transition from controlled processing to memory-based processing. This transition has traditionally been linked to the gradual withdrawal of attention (Field, 2004), making the automatic processes "uncontrollable". However, if automatic processes are uncontrollable, this cannot explain that speakers have the ability to alternate between variations in processing of the same words or sequences, for example the variation between *I am going to* and [ˈaimənə] (Bybee, 2006: 11). Rather than being an uncontrollable fusion or merging of linguistic units determining our articulation and use of a given sequence, automatization is a natural and highly useful effect of repeated use, which makes us able to communicate in accustomed and effective ways, but without removing our ability to analyze the sequences and to manipulate the units within the sequence.

The automatization of a sequence is not necessarily restricted to the process of fusion. While frequency is central in the lexicalization process for usage-based theories and models like Cognitive Grammar (Langacker 1987, 1991) and Bybee's (2001, 2006) Network Model, these theorists do not claim that fusion is the sole route to formulaicity. Actually, the claim is that the knowledge of complex structures is prior to the knowledge of single words (morphemes and phonemes), as we do not (usually) encounter language in single word utterances, but in (more or less merged) sequences of words. Instead of analysing utterances into their composite parts, and then creating formulaic sequences by fusing the lexicalized words together, we might as well store complex language structures in the first place. The frequent use of any sequence of phonemes, morphemes or words may then result in automatized units, with or without the loss of internal compositionality:

> *Repeated exposure to a particular phonological pattern (be it one we classically call a morpheme, a word, or even a sequence of words) increases speed and fluency of processing of the pattern", and, "[a]s this process is repeated, any tendency toward compositionality within the pattern is gradually reduced, leading to words and word sequences losing their compositionality if they are of high absolute or relative frequency."(Bybee and McClelland, 2005: 396).*

> *While a freshly coined expression will be fully analysable, once an expression has gained unit status through institutionalization, it is more likely that "its composite structure [...] may be activated autonomously and a gradual loss of analysability may occur" (Langacker 1987: 457, 465).*

This does not, however, entail that complex linguistic units necessarily need to be non-compositional or non-analysable. The composite parts of an automatized complex linguistic unit may still be apprehended by the language users (Tremblay, Derwing, Libben, and Westbury, 2008; Tremblay, Libben, Derwing, and Baayen, 2008). From a usage-based view, automatization includes all sorts of repeated patterns, independent of other properties of the word or sequence. Thus, compositional and fully regular multi-word sequences will be automatized if they are frequently used.

3.3.3 Conventionality

The relation between the recurrent linguistic units and patterns in language use, and the individuals' mental language systems is best captured by the notion of conventionality. Conventionality is an expression of shared knowledge, and because languages' primary

function is as communication means, the linguistic units display different degrees of conventionality: A sequence like *how do you do?* is more conventional compared to the sequence *innovative laser micromachining,* which is a term used within an industrial language register, not usually known by people that do not frequently talk about these topics.

Frequency plays an important role in this matter, since repeated exposure to a word or a sequence makes the language user able to recognize linguistic units and patterns as conventional. The usage-based view is thus that recurrent sequences in language use are probable conventional units. The use of conventional language has processing advantages both for the speaker and the hearer. The next section deals with processing within a usage-based view.

3.4 Processing – Speech Perception – Lexical Access

The usage-based language system is a dynamic system, reflecting general cognitive principles and processes. The linguistic representations are the result of processing, both perception and production, as well as categorizing and structuring processes within the system. Therefore, the processes are viewed as integral to the system, rather than an external component that serves the language system (Langacker, 1991). The consequence of this assumption is that the processing has a direct strengthening effect on the representations, which facilitates the re-selection of the representations, both in production and perception.

Speech perception is the end point in the process which starts with an acoustic signal reaching the ear and ends when the auditory input is matched with a mental representation, causing a conscious experience. Most theorists nowadays agree that perception is an interactive, and not a serial process; however, controversy exists over whether the process is sub-served by separate, autonomous systems, or by the same associative memory system.

In the subsequent sections I present two competing models of speech perception. The associative single-mechanism model, Adaptive Resonance Theory (ART), treats perception of recurrent multi-word sequences as an act of lexical access, while the generative dual-mechanism model, the Declarative/Procedural Model, assumes full analysis.

3.4.1 Adaptive Resonance Theory

Adaptive Resonance Theory (ART) is a usage-based theory of speech perception, compatible with exemplar-based theory of language storage. The theory is primarily a theory of perception, but its commitment to the usage-based thesis (cf. Section 3.2) implies that the theory also accounts for both the representation and structuring of linguistic activation patterns, because the percepts function as units of cognitive coding (Goldinger and Azuma, 2003: 309). The theory developed by Grossberg (1980) is outlined in Goldinger and Azuma (2003), and their account of the theory is summarized here.

ART describes the perception process as a cycle: first the acoustic signal activates *items* (featural clusters) in working memory. The items then activate *list chunks* in long term memory. The list chunks are the result of prior learning that may correspond to any combination of features, from phonemes to sequences of words. The items activate, through synaptic connections, the list chunks that are consistent with the input, which, in turn, receive activation in a feedback loop, creating resonance between the bottom-up signal (input) and the top-down signal (knowledge resource). The feedback loop is self-perpetuating, meaning that the bottom-up signals and the top-down signals bind into a coherent whole, even if the bottom-up signal does not perfectly match the top-down feedback. Small mismatches are ignored, but large mismatches prohibit resonance. The resulting resonance attracts attention and creates a conscious experience: *perception*.

The list chunks in ART correspond to the exemplar-based associative language system described in Section 3.3; however, the notion of psycholinguistic reality within ART is not related to the existence of activation patterns in long term memory. Here, psycholinguistic reality is related to perception. What is psycholinguistically real is a state of perceptual consequence. In normal speech situations the hearer perceives words and sequences.

A central principle within ART is masking. Larger units mask smaller units, preventing activation of a loose collection of parts. For example, upon hearing the word *jigsaw,* we perceive the entire representation, and not its smaller parts, which also are possible resonant states: *jig* and *saw*. This principle extends to more global, contextual states, which contributes to the “correct” perception of possibly ambiguous bottom-up signals.

Top-down sources may be pre-activated if a specific bottom-up signal is anticipated, which leads to accelerated resonance, despite few or distorted bottom-up sources. This anticipation may also impede resonance in cases where the stimulus is unexpected. This explains how asymmetrical support from bottom-up and top-down sources may result in correct, partly correct or wrong perception of input, as "different potential units will occasionally "win" […]" (Goldinger and Azuma, 2003: 310). Because perception is self-optimizing, strong bottom-up sources contribute to resonance, even with little top-down support. Relevant for this study is the assumption that distorted or reduced bottom-up signals (see Section 4.3.4) may still be correctly perceived if the top-down support is strong, that is, the linguistic structure is deeply entrenched.

3.4.2 The Adaptive Resonance Theory and Perception of Multi-Word Sequences

The usage-based theories hypothesize that a single associative memory system sub-serves all linguistic forms, compositional and regular as well as irregular simple and complex words and sequences. Because language use affects the mental representations, frequency effects are anticipated for all recurrent activation patterns.

The Adaptive Resonance Theory predicts that frequent multi-word sequences represent strong top-down sources, a fact that is anticipated to lead to accelerated, or easier perception, for automatized sequences, compared to the processing of non-automated sequences.

3.5 A competing View: Generative Theory

While the division between usage-based theories and modern generative theories has been reduced, some basic differences still exist. Briefly described, generative theories assume a division between lexicon and syntax. The lexicon contains the reduced, minimal units and only the irregular complex forms. This reduction is the result of the economy principle, which states that everything that can be derived by the application of a general rule can be excluded from the lexicon. This principle is an expression of a theoretical ideal which states that the simplest analysis, using fewer features and rules, is the preferred one (Crystal, 1997: 131).

3.5.1 The Declarative/Procedural Model

The Declarative/Procedural Model (DPM) of Lexicon and Grammar (Ullman, 2001) is a mental model of language storage and processing. Ullman ties these two properties of the cognitive capacity to two distinct memory systems; the lexicon of memorized words is part of a "declarative memory", while the mental grammar of rules is part of a "procedural" system. The declarative memory system is specialized for the learning and use of knowledge of facts and events, and the procedural component is presumed to be specialized for computing sequences, and hence for the learning and expression of motor and cognitive "skills" and "habits" (ibid: 37).

Ullman argues for a distinction between the two systems on two grounds. First, he links the two systems to distinct neural correlates. Different parts of the brain handle different computational operations.[13] Second, he uses psycholinguistic studies to give evidence for the distinct computations of regular and irregular morphological complex forms. In this view, regular and complex constructions do not have a unit status in memory. This lies at the heart of my inquiry, and I will in the following give a detailed description of this part.

Whereas traditional formal models have been domain specific both within the language system, and across cognitive capacities, the DPM posits a more general distribution of the cognitive capacities within two domain general systems. Ullman ties the lexical memory to the declarative memory system as they overlap in function: both lexical and conceptual knowledge are in general arbitrary. He also points to evidence from neuropsychological studies that suggest a connection between word, fact and event knowledge, this kind of information being largely sub-served by medial temporal lobe structures (Ullman, 2001: 45). The declarative memory system is thus an associative network of both word and fact knowledge, much in the same fashion as a connectionist model, however Ullman argues that the implicit knowledge of procedural skills and habits, including the mental grammar, are sub-served by a dissociated module.[14] The DPM predicts double dissociation between

[13] He presents neurolinguistic studies that support the DPM. I will comment this aspect of Ullman's study; however not in detail, because this aspect lies outside the scope of this thesis.

[14] For a comprehensive overview of "Connectionist psycholinguistics", see Christiansen and Chater, 2001).

lexicon and grammar, implicating dissociation between the more general declarative memory system and the procedural system, linked to different neural localizations.

According to the DPM the linguistic processing is not domain specific, but distributed within the two memory systems, and is thus intramorphologically modular (Ullman, 2001: 50).[15] The processing in each of the dissociated modules does not require communication between the systems, as the processes run in parallel and go on without the exchange of information between the modules. Ullman hypothesizes that the rule system receives information from the memory system during the processing of complex forms; however this information does not influence the process as such. The information is only exchanged to inhibit the execution of a rule generated form if the complex form is found through retrieval in the declarative memory system. The model does not predict that the memory system receives information from the procedural component. In this way the two systems are "largely informationally encapsulated *with respect to each other"* (ibid: 44, Ullman's italics).

The DPM is an *associative dual-mechanism model*, and Ullman contrasts it with, on one hand *traditional generative single-mechanism models,* which are both modular and domain specific, that is, the models separate the linguistic system from the other cognitive capacities. On the other hand he contrasts the DPM with *associative single-mechanism models*, including the model I use as the base for my study. I will focus on the contrasts between the DPM and associative single mechanism models, because these models share many important concerns, for example the assumption of a structured and associative lexical network. The dispute between these models is thus restricted to different views on the autonomy of syntax.

3.5.2 The Declarative/Procedural Model and Perception of Multi-Word Sequences

Ullman (2001) hypothesizes that the procedural system sub-serves at least those syntactic, morphological, and possibly also phonological computations that are fully productive, sequencing operations.[16] The declarative memory system on the other hand underlies the

[15] This extends to syntax as well, however presupposing two different computations, one symbol manipulating, and the other retrieval.

[16] Ullman (2001) focuses on morphological transformations; however he asserts that the grammatical system subserves syntactic and possibly phonological computations as well. I base my account of Ullman's model on this assertion.

transformations that are largely unproductive and that do not involve any sequencing operations. The two systems operate in parallel, meaning that both systems are working to compute the intended complex form, however if the form is found in the declarative memory, the rule-based computation is inhibited, preventing complex morphological forms as **legget* as a result of sequencing the base form *legge* ('put') and the inflectional past tense suffix *-et* for the Norwegian irregular past tense *la* ('put'), which is an irregular complex form, and hence a lexicalized item, according to Ullman.

According to Ullman then, the lexicalized items in the declarative memory do show memory effects, and Ullman points to several studies (Stemberger and MacWhinney, 1988; Prasada, Pinker and Snyder, 1990; Ullman, 1993, 1999a, all referred in Ullman, 2001) that show frequency effects for irregular complex forms. These studies show, however, no such effects for regular complex forms, and Ullman considers these studies as support for the dual-mechanism account, which claims that the regular forms are rule-products rather than stored units in the associative memory.

Applied to multi-word sequences, the DPM includes irregular sequences in the declarative memory system, and predicts that these sequences are sensitive to frequency, while the compositional and regular sequences are computed online by the procedural system. These sequences are therefore not assumed to be affected by frequency:

> *Memorized morphologically complex forms are expected to be frequency sensitive, with high-frequency forms being remembered better than low-frequency forms. Rule products that are construed from their bases in real-time should show no such "frequency effects" once one controls for access to their memorized stems, to which affixation rules are applied (Ullman, 2001: 52)*

3.6 The Competing Assumptions Regarding Perception of Frequent Recurrent Sequences

The associative single-mechanism models assert that a single associative memory subsumes the learning, the representation and the computation of both linguistic and other conceptual knowledge. Within this system, there is no categorical distinction between non-compositional and compositional forms. There are no abstract rules and thus no distinct system that manipulates symbols by these rules (Ullman, 2001: 49). These models predict

that both regular and irregular complex linguistic forms are computed by the same associative system, whereas the DPM predicts that the different transformations will posit psychological dissociations, in which the regular forms depend on the rule system, while the irregular forms depend on the associative memory system. On these grounds, the DPM predicts that the complex linguistic forms computed by the associative memory will show memory effects, whereas the compositional forms computed via the procedural system will not. On the other hand, the associative single-mechanism models predict memory effects for all lexicalized forms.

The different theoretical assumptions of what belongs in memory and what does not, described in the previous sections, generate contradictory hypotheses about how fully compositional multi-word sequences are computed. The exemplar model of storage and processing assumes that compositional, frequent word sequences are memorized by the language user. The strong representation in memory is seen as a consequence of the sequences' frequency in use: what we hear and use often is remembered, whether it is smaller or larger constructions. The strong representation in memory contributes to ease both production and perception of the sequences.

This assumption may seem intuitive, however models that operate with a distinction between lexicon and syntax, like the Declarative/Procedural Model (Ullman, 2001), assume that the processing of sequences are only affected by the frequency of the sequences' composite parts. Therefore, sequences containing high frequency words are processed faster and more easily than sequences containing low frequency words. So, as long as the sequences' elements are controlled for frequency, the frequency of the whole sequence is viewed as an irrelevant artefact in the processing of compositional and regular multi-word sequences (Ullman, 2001: 52).

As the present thesis is founded in usage-based theory, I aim to find support for the hypothesis that recurrent multi-word sequences have mental representation as entrenched activation patterns in language users' minds. This assumption was also the starting point for the study described in Section 2.3.3 (Schmitt et al., 2004). Schmitt et al. did however conclude that a sequence's frequency does not predict its status as a formulaic sequence, which may be seen as supporting Ullman's assumptions. The next chapter presents a psycholinguistic experiment, which challenges these assumptions.

4 The Psycholinguistic Experiment

4.1 Introduction

The Psycholinguistic experiment is designed to test the assumption that fully regular, recurrent multi-word sequences are entrenched memory units. This hypothesis is consistent with a usage-based view of language storage and processing, and it is contrasted with the null hypothesis derived from a generative dual-mechanism model, which assumes that these sequences are online computations subsumed by an abstract rule-system.

The holistic representation of multi-word sequences has been shown to contribute to ease the perception of these sequences (cf. Section 2.3.2). There are no ways to observe this process directly; however a psycholinguistic experiment which aims at investigating whether frequent recurrent sequences are perceived more easily than supposedly non-automatized sequences may give us some insight into this matter. If a difference in perception of the two frequency groups is found, this may indicate that frequency does affect the mental representations of larger constructions, like it does for single words. It may also indicate that frequency of use affects sequences in general, regardless of the absence of properties associated with formulaicity. Evidence of holistic representation of fully compositional sequences questions the classification criteria used in the research field of formulaicity as well as the basic assumptions regarding language representation in mind, described in Chapter 2.

Within usage-based theories, the recurrent sequences are assumed to be entrenched activation patterns. This represents a top-down support in the perception process, while the speech signal represents a bottom-up source. According to Adaptive Resonance Theory (ART) (Goldinger and Azuma, 2003, described in Section 3.4.1), reduced bottom-up speech signals demand a stronger top-down support for the utterance to be correctly perceived by the hearer. To test this assumption, the psycholinguistic experiment consists of a set of frequent target sequences, which is tested against a set of infrequent target sequences. Both sets are taped and masked with noise to reduce the acoustic signal. According to the ART model, expected results of the masking will be that the test subjects perceive either 1) the correct

sequence, if the top-down source is strong, 2) a wrong sequence, if the bottom-up and the top-down source coalesce in an approximation because the top-down source is missing, or is weak, 3) fragments of the sequence, or 4) unintelligible speech signals.

The advantage of psycholinguistic experiments is that they offer insight into language processes. However, as I point out in Section 5.2 below, experiments must be carefully designed to avoid validity problems. The present test is a controlled experiment, reducing the external factors to be accounted for compared to experiments conducted in natural environments. It is rigidly designed to reduce the extraneous variables that affect the results. By isolating the sequences from their natural context it is possible to avoid that the participants direct their attention to the larger context (see Section 3.3.1 about attention).

The assumption I am testing is that high frequency contributes to a strong mental representation, which *eases* the perception of the sequences. I operationalize "ease" in terms of statistical behaviour: The participants are hypothesized to be able to repeat more of the frequent recurrent sequences than the infrequent sequences. The recurrent multiword sequences in a corpus and the entrenched memory units in mind are assumed to correlate. The fact that usage-based theories assume that the relation is bidirectional, so that frequent use leads to strong mental representation and a strong mental representation increases the probability that the construction will be used, may seem to be like the classical problem: what came first – the chicken or the egg? However, the bi-directionality does not change the assumption that use strengthens the mental representation. The relation between frequency of use and automatization is actually not disputed – as long as we are talking of single words. However, I intend to find support for this kind of correlation also for sequences of words.

A possible disadvantage of controlled experiments is that the test situation produces results that may not extend to real-life situations: We do not usually perceive the target sequences in isolation. In natural settings, contextual information is present and will contribute to successful perception (Field, 2004). Also, the speech signals and possible noise usually have separate and physically distant sources, and do not reach the ears as a merged signal, which the signals in the present experiment actually do. It is also likely that speakers use compensatory strategies to ease listeners' perception of less conventional language use, by articulating more carefully - or just talking more loudly. Thus, infrequent non-automated language is not necessarily harder to perceive than frequent and automatized language in

normal speech situations. Despite the fact that controlled experiments do not test variables in their natural contexts, in this case, it may be just the reason for choosing an experimental design over collecting data from real-life situations. A hypothesized difference in mental representation may be more accessible in a controlled experiment because we have the opportunity to remove all the extraneous variables, that is, the effects of the compensatory strategies mentioned above. A possible negative consequence of the manipulated situation is that even highly entrenched sequences may fail to be activated. In defence of the present experiment, I argue that it is relatively naturalistic. The experiment simulates real-life situations where speech signals are distorted by noise. This is by far the normal situation, as we are surrounded by noise of different sources more often than we converse in totally silent surroundings.

Ullman (2001) predicts no difference in perception between frequent and infrequent sequences, as long as both sets of target sequences are balanced for word frequency (cf. Section 3.5.1). This represents a null-hypothesis, and the competing hypothesis for my psycholinguistic experiment.

4.2 Material

The test sequences consist of recurrent sequences extracted from the Norwegian oral corpus NoTa-Oslo. This is a corpus of interviews and conversations gathered from 166 informants in the age between 17 and 60, born and raised in or in the suburbs of Oslo. The informants are balanced according to age, sex, residence and education. The corpus consists of approximately 900 000 words, which are orthographically transcribed and morphologically tagged. It is searchable in a specialized interface and the transcriptions are linked to both sound and video files. The corpus is a large collection of text which represents a particular register that may be termed “conversational language”. I assume that this specific register is known and used by practically all east-Norwegian speakers. A register is recognized by its predictable reoccurring patterns of language use, and it is the frequent reoccurring patterns in NoTa-Oslo which are of specific interest for this study.

The use of frequency data collected from corpora in psycholinguistic approaches to grammar is common practice (Merlo and Stevenson, 2002). However, the assumption that corpus-

derived patterns are representative of a language user's knowledge of language is still disputed. An argument against the use of corpus-derived statistical material is that corpora are not usually balanced according to register, and are therefore over-represented with regard to certain types of language usage and thus under-represented for others. As a consequence, the statistical data will reflect this imbalance, and therefore, "statistical studies based on corpus should be taken with many large pinches of salt" (Johannessen 2003: 164, my translation). There is indeed reason to be cautious when inferring from corpus-derived statistical data to mental representations. Nevertheless, if the corpus is representative of a defined sub-part, or a specific register, of a given language, the statistically derived data may very well reflect the mental representations of the language users that master and use this register. With these considerations in mind, I argue that the NoTa-Oslo corpus is balanced because it is restricted to spoken Norwegian, and to a specific register. The register reflects conventional knowledge of how, and with what means, to converse in east-Norwegian. Since this kind of knowledge is rather basic, I assume that the recurrent sequences in NoTa-Oslo are conventional knowledge.

The N-grams of the NoTa-Oslo corpus were identified using the Ngram Statistics Package (NSP).[17] NSP is software used for analyzing n-grams in text files. An n-gram is the sub-sequence of *n*, that is, any number of items (i.e. words) from a given sequence. The NSP consists of the software program "count.pl", which was fed the flat text files from NoTa-Oslo as input. It then generated a list of all the 2- to 5-grams that occurred in the file. The n-grams and their frequencies were listed in descending order of their frequency. An example of n-grams from a text is given below:

When I was born, I was so surprised I didn't talk for a year and a half [18]

The text generates these token words:

When - I - was - born - I - was - so - surprised - I - did - not - talk - for - a - year - and - a - half

[17] Available online at: <http://ngram.sourceforge.net/>

[18] Quote by Gracie Allen

Then, the following bi-grams, formed by contiguous tokens, can be identified:

1)when<>I 2)I<>was 3)was<>born 4)born<>I 5)I<>was 6) was<>so 7) so<>surprised 8)surprised<>I 9)I<>did 10)did<>not 11) not<>talk 12)talk<>for 13)for<>a 14)a<>year 15)year<>and 16)and<>a 17)a<>half

Of these 17 bi-grams, the bi-gram I<>was occurs two times, while the remainder occur only once. The 4-grams identified in this text are:

1)when<>I<>was<>born 2)I<>was<>born<>I 3)was<>born<>I<>was
4)born<>I<>was<>so 5)I<>was<>so<>surprised 6)was<>so<>surprised<>I
7)so<>surprised<>I<>did 8)surprised<>I<>did<>not 9)I<>did<>not<>talk
10)did<>not<>talk<>for 11)not<>talk<>for<>a 12)talk<>for<>a<>year
13)for<>a<>year<>and 14)a<>year<>and<>a 15)year<>and<>a<>half

The hundred most frequent 4- and 5-grams from NoTa-Oslo were chosen as the basis for further selection for the frequent targets, while a selection of the hundred most frequent 2- and 3-grams were used to construct the infrequent target sequences in the experiment.

4.2.1 Properties of the Recurrent Sequences from NoTa-Oslo

When you extract 4- and 5-grams from the NoTa-Oslo corpus, you get a vast number of sequences that range from 1 occurrence, for example *får meg lappen og bil* ('get myself certificate and car') or *mye mer heftig performanceart enn* ('much more intense performance art than') to the most frequent sequences, for example *for å si det sånn* ('so to say'), which occurs 116 times, *ja ja ja ja ja* ('yes yes yes yes yes'), which occurs 112 times, and *ja men det er jo* ('yes but after all'), which occurs 68 times in the corpus. The difference between the frequent and the infrequent sequences in NoTa-Oslo is that a majority of the infrequent sequences consist of infrequent and salient words, while the frequent sequences consist of frequent, "semantically light", and more grammatical words. What characterizes the frequent sequences is that they call forth recognition. Upon hearing these sequences, people say that they recognize them as familiar. According to Bybee (2006), this is clear evidence that the sequences are conventional.

The reasons for selecting mainly 5-grams, in addition to a few 4-grams, are part deliberate and part incidental. The NoTa-Oslo corpus was in advance of the present study already analysed into 2- to 5-grams, and I used this analysis as the starting point for my investigation. I could have requested larger sequences; however, the restriction in size of the sequences does not pose any restrictions for the study. The properties of the material are in fact rather optimal for the purpose of the study. Sequences larger than five words are not usually represented more than once in relatively small corpora, like NoTa-Oslo, while the range of five-word sequences is large enough to be interesting. A disadvantage of larger sequences in this specific experimental design is that such sequences tend to be more idiomatic, a property that firmly places the sequence in the lexicon, also within generative models. This is not the case for the four- and five-word sequences. These sequences are regarded as fully compositional, and therefore, they do not qualify as formulaic sequences by the established criteria defining formulaicity within the research field (cf. Section 2.3).

4.2.2 The Material Represented in an Associative Network

The hundred most frequent 5-grams occur between 14 and 116 times in the corpus, and many of the sequences are partly overlapping, like for instance *og så var det en/og så er det en/og så var det jo* (and then was it a /and then is there a /and then were there yes). This may be represented in an associative network (cf. Section 3.3.2), illustrated in Figure 5 below.

The network consists of selected sequences from the hundred most frequent sequences from NoTa-Oslo. Sequences that represent false starts or which are the repetition of one single word, like the sequence *ja ja ja ja ja,* are not included in the network, because they would make the network too complex. This would preclude the point that some of the sequences are part of tighter networks of formally similar sequences, represented in the illustration as lines between similar words, while others are more fixated, represented by the more remotely located sequences with few nodes attached to them. The associations between the sequences in Figure 5 are based on formal properties; however, I have attempted to place semantically related sequences in the same areas. The sequences are separated into five frequency groups, represented in the network by different thickness of the boxes:

Figure 5 An associative network of frequent 5-grams from No-Ta-Oslo

Characteristic of the recurrent sequences is that most of the instances are unaccented and phonologically reduced (cf. Section 3.3.2). According to Bybee (2001), this reduction indicates that the sequences are processed as routines, because words that frequently occur together "begin to behave phonologically as if they constituted a single word" (Bybee, 2001: 161). The sequences are losing their internal complexity, which indicates that speakers treat these sequences as units of production and that hearers perceive the sequences as known or conventional units. Another indication that the recurrent sequences are treated as units is the occasional use on the Internet of the sequences *for å si det sånn* ('so to speak') and *holdt jeg på å si* ('I almost said'), written in one connected string: *Foråsidesånn* and *holdtjegpååsi.*

A property of specific relevance for the psycholinguistic experiment is that the sequences are *fully compositional* (cf. Section 1.5). They are "word sequences that are conventionalized, but predictable in all other ways" (Bybee, 2006: 4). The fact that the material consists of relatively short, self-contained and compositional sequences, makes it ideal for the psycholinguistic test, as the properties of the sequences do not place them firmly on either side of the traditional lexicalized/non-lexicalized division of language representation. The sequences are frequent and compositional, which leads to the question: are they perceived with ease, as language routines, or are they perceived bottom up, morpheme by morpheme?

4.3 Test-design

The psycholinguistic experiment is intended to test if a difference in frequency creates a difference in perception; therefore the set of frequent recurrent sequences is tested against a set of sequences that are infrequent and supposedly non-automatized. The set of infrequent four- and five-word sequences consists of the same high frequency words as the set of frequent recurrent sequences. This is done to avoid a difference in perception caused by differences in word frequencies. A third set of sequences consists of dummy sentences to prevent, or at least reduce priming, and also to reduce practice and fatigue effects (see Section 4.3.3). The test sequences are taped and masked with noise, which makes the input signals reduced and harder to perceive. The masking is done to force the participants to rely more on top-down knowledge in perception, in lack of complete or clear bottom-up speech signals.

4.3.1 Selection of the Frequent Target Sequences

A selection of 29 frequent test sequences was made out of the hundred most frequent four- and five-word sequences extracted from the NoTa-Oslo corpus. The selection criteria which were used are as follows: The most frequent sequences were chosen. In cases where two or more sequences overlap by three or four words, the more frequent sequence was selected, or if the less frequent sequence is a more self-contained whole, this sequence was preferred instead. An example is the choice of one of several overlapping sequences (test sequence nr. **11** in bold (see appendix I): *det er jo ikke noe/* ***det er jo ikke det/*** *men det er jo ikke/ det er jo ikke så/ det var jo ikke noe* (it is yes not any/ it is yes not it/ but it is yes not/ it is yes not so/ it was yes not any). This sequence is the second most frequent, but is preferred over the more frequent sequence because it may be used in isolation, while the other sequences are always part of larger intonation units.

Some sequences were excluded from the list since they clearly are the result of the collection techniques used for the NoTa-Oslo project. Both conversation and interviews were used, and some of the five- and four-word sequences were probable responses to interview questions, e.g. "Where were you born and raised?", resulting in the frequent response: *Jeg er født og oppvokst…* ('I am born and raised…'). This does not exclude the possibility that these constructions in fact are memory-units. Nonetheless, their frequency in this particular corpus is probably much higher relative to what is expected for a pure conversational corpus, while the remainder of the constructions are also expected to have relatively high frequency in equivalent or similar corpora, for example the Norwegian Big Brother corpus.[19] All of the selected frequent target sequences are represented one or more times in the material as self-contained wholes. The instantiations are represented with a pause either before or after the unit, or both. The final set of frequent target sequences with their frequency in the NoTa-Oslo corpus and English translations are shown in Table 2 below:

[19] BigBrother-korpuset, Tekstlaboratoriet, ILN, Universitetet i Oslo. Online available at: <http://www.tekstlab.uio.no/nota/bigbrother/>

Table 2 Frequent target sequences

Seq. Nr.	Frequent sequences	Freq. in NoTa-Oslo	Word-by-word translation	English equivalents
1.	ja men det er jo	68	Yes but it is yes	'Yes but it is'
2.	for å si det sånn	116	For to say it there	'So to speak'
3.	det er det som er	61	It is that that is	'That's what it is'
4.	nei jeg vet ikke jeg	52	No I know not I	'No I do not know'
5.	det er i hvert fall	52	It is in every case	'It is in any case'
6.	jeg tror ikke det er	29	I believe not it is	'I do not think it is'
7.	et eller annet sånt noe	19	One or other there something	'Something like that'
8.	det er ikke så veldig	39	It is not so very	'It is not that/quite'
9.	jeg har lyst til å	19	I have desire till to	'I would like to'
10.	og så er det jo	25	And so is it yes	'And then there is'
11.	det er jo ikke det	34	It is yes not it	'It is not'
12.	i og med at jeg	17	In and with that I	'Because I'
13.	det er klart det er	29	It is clear it is	'That is clear'
14.	jeg er ikke helt sikker	16	I an not quite sure	'I am not quite sure'
15.	det er på en måte	19	It is on one/a way	'It is in a way'
16.	og da var det jo	18	And then was it yes	'And then there was'
17.	det er ikke noe problem	15	It is not any problem	'That is no problem'
18.	nei jeg tror ikke det	17	No I believe not it	'No I do not believe it is'
19.	jeg er veldig glad i	17	I am very fond/glad in	'I am vary fond of'
20.	holdt jeg på å si	47	Held I on to say	'I almost said'
21.	det er det det er	27	It is that it is	'That's how it is'
22.	ikke så veldig mye	68	Not so very much	'Not that much'
23.	ja det er jo det	32	Yes it is yes that	'Yes it is'
24.	det er klart det	63	It is clear that	'Of course'
25.	og det syns jeg er	17	And that think I is	'And I believe that'
26.	det er ikke noe sånn	32	It is not any problem	'It is nothing like that'
27.	ja det er sant	53	Yes it is true	'Yes that's true'
28.	men det er jo det	26	But it is yes it	'But it is'
29.	jeg syns det er veldig	19	I think it is very	'I think it's very/really'

4.3.2 Construction of the Infrequent Target Sequences

The infrequent target sequences had to fulfil two demands: Firstly, they had to be infrequent and supposedly non-automatized sequences of the same size as the frequent targets. Secondly, their component words had to be within the same frequency range as the component words of the frequent target sequences. This is due to Ullman's demand that the composite words of the sequences must be balanced for frequency to prevent processing advantages for sequences that consist of elements of high frequency over sequences

consisting of low-frequency elements (cf. Section 3.5.2). Two- and three-word combinations extracted from the NoTa-Oslo corpus, which are also part of the frequent four- and five-word target sequences were combined into unusual five-word sequences with as few as eleven or fewer tokens in searches on www.google.no. These sequences are supposedly non-automated, and thus not assumed to be conventional units. The method was used to avoid test sequences that were salient or actually frequently used, which is the case for the low frequency five-word sequences from NoTa-Oslo. Most of the sequences are characterized by hesitations or false starts, or they consist of low frequency items that would be far more salient than the elements in the high frequency sequences. This is undesirable as these extraneous variables certainly would have affected the results. Even sequences that occur in the corpus only twice are problematic, because searches on www.google.no show that these low frequency sequences are of relatively high frequency on the Internet. This fact is due to the small size of the NoTa-Oslo corpus.

The final set of infrequent target sequences with their frequency on Internet and English translations are shown in Table 3 below:[20] The sequences are separated by the symbol "+", which marks the joint position between the 2- and 3-grams.

[20] The table displays frequency of occurrences on www.google.no 2009.05.05.

Table 3 The infrequent target sequences consisting of 3+2 word sequences extracted from NoTa-Oslo.

Seq. Nr.	Infrequent sequences	Freq. on Internet	Word-by-word translation	English equivalents
30.	det er sånn+ jeg syns	2	It is that I think	It's that that I think
31.	og det var +hvis du	2	And that was if you	And that was if you
32.	men det var+ jeg bare	2	But that was I just	But that's just how I were
33.	det er det+ å da	0	It is that to then	It is just to
34.	at det er +at jeg	7	That it is that I	That it is that I
35.	på en måte+ er sånn	1	On one/a way is that	In a way is such
36.	det er så+ jeg ikke	2	It is so I not	It's like I don't
37.	det var en +jeg skal	0	It was one I shall	There was one I was
38.	nei det var+ på det	0	No it was on that	No that's on the
39.	er ikke noe+ på det	3	Is not some on that	Is nothing on
40.	jeg tror det+ jeg var	1	I think it I was	I think what I was
41.	nå er det +som er	5	Now is that which is	Now is what it is
42.	er jo det+ hvis du	1	Is yes that if you	is if you
43.	det har jeg+ som er	0	That have I which is	That's what I have which is
44.	så var det+ jeg har	2	So was it I have	Then was what I have
45.	at det var +hvis du	0	That it was if you	That it was if you
46.	så har jeg+ en gang	5	So have I one time	then once I have
47.	ja og så+ er ikke	0	Yes and so is not	Yes and so it is not
48.	for det er +jeg bare	0	For that is I just	For that is just hove I am
49.	er jo ikke+ på en	0	Is yes not o none/a	Is not on a/one
50.	og så er+ så mye	0	And so is so much	And then is so much
51.	jeg har ikke+ det at	1	I have not that is	I have not that
52.	det er liksom+ at jeg	5	It is just that I	It is just like that I
53.	er jo det+ når jeg	0	Is yes that when I	Is when I
54.	i hvert fall+ er sånn	2	In each case is that	In any case is that
55.	er det ikke+ ja og	0	Is it not yes and	Is it not yes and
56.	vet ikke jeg +at det	5	Know not I that it	I do not know that it
57.	er det jo+ til å	11	Is it yes till to	It is to
58.	da var det +jeg hadde	2	Then was it I had	Then what I had was
59.	det er jo+ og det	7	It is yes and that	It is after all

The infrequent target sequences were rated as more or less acceptable by three Norwegian language users - two linguistics students and one computer scientist. Not surprisingly, the two linguists were more liberal than the non-linguist. The infrequent sequences were first presented in isolation, and the result was that 3 of 30, 2 of 30 and 9 of 30 sequences were viewed as non-acceptable for the three informants respectively. Next, the sequences were presented in context, something that lead to a reduction in unacceptable sequences, now 0 of 30, 2 of 30 and 4 of 30 were rated as strange or unnatural. The sequences that were

considered unnatural were viewed as fully intelligible, however the informants described them as unfamiliar and deviating from orthographic conventions.

4.3.3 Dummy Sequences

Since the same elements are used in several of the target sequences and across the two sets, there was need for something that, if not prevented, at least reduced priming. Priming is not considered to be a major problem for this test, as the perception of one or a few elements does not guarantee the perception of the whole sequence (cf. Section 3.4.2). Only correct reproduction of the complete sequence is scored. I still chose to use dummy sequences to minimise the possibility that one sequence primed the subsequent one in cases where the sequences were partly overlapping or similar in some manner. Another advantage of using dummies is that it prevents the targets from being "wasted", as the participants use some time to adapt to the task, a condition termed "practice effects". For this reason, it is advantageous to start the test with sequences that get the subjects onto the right track and later exclude these practice sequences from the test.

The dummy sequences are mainly taken from the table of contents of two books.[21] The sequences are fragments ranging from two to six words, for the most part low-frequency words from different registers than the target sequences. The intention was to use dummies that differed drastically from the targets so they would not prime the targets, and also made the task less predictable for the participants. In addition to these index sequences, five practice sequences of the same type and quality as the frequent targets were presented in the beginning of the test to activate the "conversational register". The final set of dummy sequences amounted to 65; six at the beginning of the test, and one in between each target sequence (see Appendix I).

[21]"Skrive for å lære" (Dysthe et al., 2000), and "Retorikk: De viktigste retoriske figurer belyst ved eksempler fra riksmålets litteratur" (Nordahl, 1994).

4.3.4 White Noise

Besides developing suitable test sequences, a method to degrade the input signal was needed. In order to reduce the intelligibility of the speech signals, white noise was added. White noise is "a constant hiss like an extended fricative or like radio static. It is *aperiodic:* its average loudness remains constant over the entire range of audible frequencies" (Field, 2004, his italics). The noise is continuous and "flat", not moulded to fit the speech signal, but constant at the same frequency and intensity. The specific noise used for the psycholinguistic experiment is a 44 kHz, 32 bit float white noise.

4.3.5 Developing the Instrument

The total of 122 sequences was recorded with the sound program Audacity.[22] The sequences were read in a naturalistic way, reflecting their normal pronunciation in context. The typical pronunciation pattern for the frequent target sequences is found in the sound files linked to the orthographical transcriptions in the NoTa-Oslo corpus, whereas the pronunciation of the infrequent target sequences attempts to match the natural pronunciation of the sequences inserted in their context (cf. Section 4.3.3). A pilot test was done to find out if any of the sequences were perceived as being louder than the rest, which was the case, and they were therefore checked for intensity and pitch in the sound program Praat.[23] The sequences that were either below or above the range of 72 – 76 dB at the highest intensity, were recorded a second time to make sure that all the target sequences were scaled within this range. Listening to the scaled sequences without noise gave after this adjustment a sensation of 'sameness'. The average highest dB for the frequent sequences is 74.19dB. The highest dB within this set is 75.94dB, and the lowest 72.48dB. The average highest dB for the infrequent sequences is 74.07dB. The highest dB within the set is 75.94dB, and the lowest 72.2dB. This is approximately equal values for both sets, and the marginal variation is not assumed to affect the results. Regarding pitch, the average value for the frequent sequences is 278.54Hz, ranging from 478.7Hz to 232.4Hz. The average value for the infrequent sequences is 273.18Hz, ranging from 490.6Hz to 228.2Hz. The variation between the two sets of target

[22] Free software program, downloaded at <http://audacity.sourceforge.net/>

[23] Free software program, downloaded at <http://www.fon.hum.uva.nl/praat/>

sequences is so small that it is not assumed to affect the results. The two sets of target sequences consist of approximately the same words, something that prevents large variation in perceptual prominence between the sets. Variation in intensity, pitch and perceptual prominence are indeed extraneous variables that may affect the results; however the combination of equalizing these factors between the two sets and also quantifying the number of sequences in the test, that is, the greater number of test sequences, makes it less probable that the variation between the sequences will have a decisive impact on the results.

The level of noise masking the test sequences was established after a second pilot. The goal was a test battery that would achieve about 80 % correct reproductions of the frequent target sequences. The grade of difficulty of the test is important to control, to avoid what is called a "ceiling effect", which means that the majority of scores are at or near the maximum possible for the test or measurement, and which in this specific case would obstruct the desired top-down processing effect. Thus, if the noise level is too low, it would be hard to achieve any possible statistically significant difference in reproduction of the target sequences. It is equally important to avoid the target sequences being too difficult to comprehend. It is psychologically undesirable to give tests with an expected score as low as approximately 30%, because this may cause the participant to feel inadequacy (Myhrum, personal communication). Ideally the score should be no less than 50%. The noise level was set to 5dB for the target sequences, which gave a 80 % score for the frequent target sequences in the second pilot. For the dummy sequences, the noise level was set to 3dB. The reason for choosing a lower noise level for the dummies was that these sequences are more difficult to comprehend than the targets. The dummies were intentionally still difficult to perceive, to prevent a too noticeable difference compared to the target sentences, which might bring the attention away from the task. The outcome was that many of the salient words in the dummy sequences were perceived, giving the participants the sensation of mastering the task. The participants were also made aware that many of the sequences might be difficult or impossible to perceive, hence they should not despair if they could not repeat one or several sequences in a row.

A windows program, SpeechUtil, which coordinated the sound files with a noise file, was used to present the sequences masked with noise.[24] The noise starts off 0.5 seconds before each sound file, and ends at the same point as the sound file. The result gives the impression of a buzzing noise with speech "underneath". The program gives the opportunity to present the speech files at two different modi; either all the speech files successively with a short fixated break between each file, or as individual speech files. The latter modus was chosen to give the test participants time to repeat each sequence at their own pace, with the restriction that they must give immediate response without unnecessary delay. Another advantage of the self-regulated test speed is that by adjusting the speed to the individual participant, the faster ones did not have to wait several seconds for the next sequence to appear which might have caused a sensation of boredom or impatience.

4.3.6 Participants

The variable most difficult to control is the individual participant. To be able to make a statistical analysis, I have included 32 participants in the experiment.

The 32 participants consisted of native Norwegian speakers, within the age span from 21 to 58 years (22 female and 8 male). All the participants speak an east Norwegian dialect, which is an approximation to the dialect or dialects in Oslo. The participants were all ignorant of the aim of the experiment and the theme of the thesis. They were mostly recruited within the university community in Oslo, while a few were recruited through friends and relatives. A note asking for participants was posted several places on campus, something that generated only two subjects, so most of the participants were contacted personally and asked if they had the opportunity to participate in a psycholinguistic test. The participants were not offered any reward for their contribution; however they were eager to do their best. I suspect this is a psychological effect of hearing the word "test".

Compared to the informants for the NoTa-Oslo corpus, the participants' profiles are not controlled according to birth place and where they have been raised, only according to age and dialect. Most of the participants have higher education, while the informants for NoTa-

[24]SpeechUtil, version 5, developed by Dan Freed, House Ear Institute, June 2004.

Oslo included persons with both higher and lower levels of education. Even though the participants are not balanced for these factors, I have reason to believe that they are representative regarding knowledge and use of the specific register under investigation. As mentioned earlier (see Section 4.2.1), the knowledge of this register may indeed vary both between language users and between groups of language users. Nonetheless, the frequent recurrent sequences are conventional and, if not used, at least known by all, or nearly all, east Norwegian speakers, and thus, the test participants.

The participants made a self report on their hearing condition, and none reported any severe hearing deficiency. This method does not of course exclude the possibility that some of the participants in fact do have hearing deficiencies. Nonetheless, if this actually is the case, it is not supposed to be a major disadvantage, because reduced hearing should affect both target sets in the same way, leading to the same hypothesized difference between the sets.

4.3.7 Procedure

The testing was in most cases carried out in quiet surroundings at the University of Oslo. I had a personal office at my disposal with a comfortable chair for the participant. The test equipment was placed in a fixed position, securing a homogeneous situational context. Seven of the participants were tested elsewhere; however the different arrangements were attempted to be as identical as possible. The stimuli were kept constant irrespective of the changing test contexts. The test equipment consisted of a laptop computer, headphones, and a digital recorder.

The participants were tested one by one. Before the test started, each participant read a test instruction where the task was briefly outlined. The instruction included a short description of the test's goal. This was done to give the subject an impression of what she or he could expect; however, without revealing the actual purpose of the test. The participants were asked to repeat the sequences as similarly as possible to what they heard. They were asked to give an immediate response, without any unnecessary delay. If they did not perceive the whole sequence, or if they were uncertain of what they had heard, they were encouraged to make a guess. They were allowed to ask explanatory questions if something was unclear, as long as it did not concern the purpose of the test. The participants were equipped with headphones and seated facing the recording equipment. The testing was executed by me

only, controlling the playback speed and the recorder. The participants were not able to listen to each target sequence more than once.

The psycholinguistic test is relatively short, taking approximately fifteen minutes from start to finish. Still, it seemed to be the case that several of the test subjects were experiencing fatigue effects, that is, the sensation of boredom or loss of attention towards the end of the test. Another expected effect was the practice effect, which affects the results as the subjects get better at solving the task the more they get accustomed to the procedure. To be sure those effects were avoided, the second half of the participants were presented with the test sequences in the opposite order. Only the first six dummy sequences kept the same position at the beginning of the test to avoid losing the targets because of practice effects.

The final data set consisted of the response from 30 of the 32 participants, and the results for 29 of the 30 frequent target sequences. The seventeen first participants were tested with the sequences presented in order from 1 – 122, while the last fifteen participants were presented with the sequences in the opposite order. Two of the participants from the first group were excluded from the test. The choice was rather easy to make, as the recorder broke down during the testing of the first participant, and I was not noting the scores during this specific test session. The result was that half the test was missing. In addition, I adjusted the noise for the dummy sequences after this first participant. A second participant was excluded because the results diverged rather a lot from the mean. The participant scored 11 out of 30 frequent sequences, while none of the infrequent sequences were correctly reproduced.

When I was going through the results, I found that I had made a slip: Sequence nr. 20 (see Section 4.3.1) appeared two times in the test. I decided to exclude the results for the sequence that received the highest score from the statistical analysis. An possible bias will therefore be against my own hypothesis.

A two-part scoring system was used to quantify the participants' performance. The response was divided into two categories. The exact, correct reproduction of a sequence, without any missing or additional elements, was scored 1, all other responses, or missing responses were scored 0. For each target sequence, the number of participants that scored 1 was noted. For the qualitative analysis, all attempts at reproducing the sequences were orthographically

transcribed to see if the actual response, if not correctly reproduced, could give insight in the way reduced input is processed.

4.4 Results

The results of the experiment are dealt with in two main ways, statistical and qualitative. The statistic analysis includes the total scores for the two frequency groups, the performance for each participant and the score for each target sequence. I have provided a two-tailed fisher's exact test to find the P value for the association between the two frequency groups overall, and for the response for each participant.[25] This statistic analysis contributes to give a clear picture of how the sequences are distributed, and makes a good basis for further analysis. The qualitative results of the experiment are presented as an orthographic transcription of the incorrect reproduction of both the frequent and the infrequent targets (see appendix III). The qualitative data are used to investigate the actual, incorrect response from both frequency groups to see if there are any tendencies, and if this can be explained in reference to frequency.

4.4.1 Quantitative Results

Overall, the participants correctly reproduced 598 of the 870[26] frequent target sequences, while they correctly reproduced 196 of the 900 infrequent target sequences, illustrated in Figure 6 below.

Correct response includes only the exact reproduction of the stimuli, all other responses scored 0, as mentioned above. The results show that the frequent target sequences were correctly reproduced more than three times as often as the infrequent targets. With the Fisher's exact test, the association between the total score for each target group gives a two-tailed P value of less than 0.0001, which is considered to be "extremely [*sic*] statistically

[25] Graphpad is a statistical tool, available online at <http://www.graphpad.com>

[26] The initial number was 900; however, one of the frequent targets was erroneously duplicated in the test, and was therefore eliminated from the data set (see Section 4.3.7)

significant" (http://www.graphpad.com).[27] The results indicate that this considerable difference in reproduction between the frequent and the infrequent target sequences is a consequence of difference in frequency, and not a result of variation among participants.

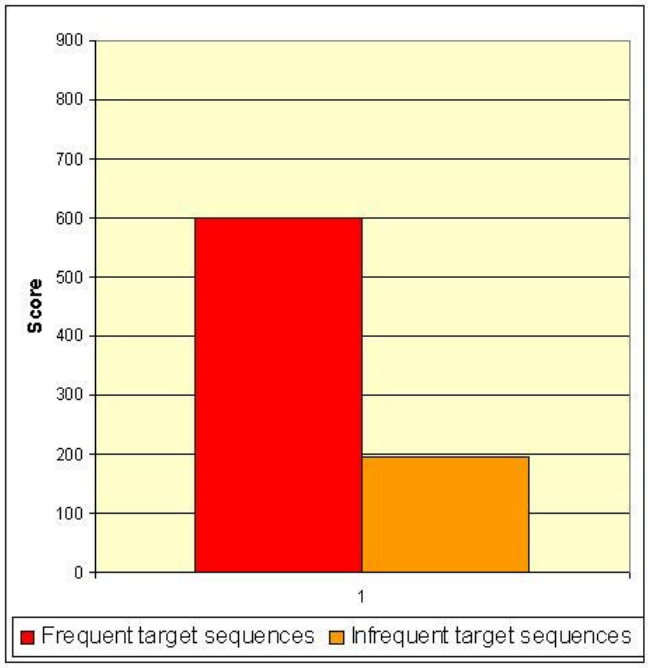


Figure 6 Overall score for the frequent and the infrequent target sequences

The participants scored 598 out of 870 possible for the frequent target sequences, which is 68.7 % correct reproduction and thus lower than anticipated (cf. Section 4.3.5). For the infrequent sequences, the total score is 20.8 %. A likely reason for the low score for the frequent target sequences is that they are more difficult to perceive when they are inserted into the test together with the infrequent and the dummy sequences than they are when they are presented in isolation. The participants are forced to exert themselves to comprehend the more difficult parts of the test, which I suspect will affect their overall achievement. A reduction in noise, so that the score for the frequent target sequences would have been closer to the anticipated 80 % score, would presumably also have lead to a higher score for the

[27] This is actually how the online statistical tool Graphpad (see note 25) describes the statistical relation between the total scores for the two target groups.

infrequent target sequences. This anticipated correlation between the two frequency groups is in fact confirmed in the result for each participant, as illustrated in Figure 7 below.

Participants' scores

The scores and P-values for each participant are given in Table 4 below. The number of target sequences included in the experiment was 29 frequent target sequences and 30 infrequent target sequences. The mean performance for the frequent target sequences is 19.9 with a standard deviation of 3.9. For the infrequent targets, the mean performance is 6.5 with a standard deviation of 3.2. All the participants scored relatively high on the frequent targets compared to the scores for the infrequent sequences. With a significance level at 0.05, 19 out of 30 participants produced a significant difference between the scores for the two frequency groups, as illustrated with bold numbers in Table 4, while 11 of the participants did not correctly reproduce a significantly higher number of frequent target sequences compared to the infrequent targets. Nonetheless, the participants that produced the least difference between the two frequency groups still reproduced correctly nearly twice as many frequent targets as infrequent targets. Participant number 19 and number 26 both correctly reproduced 20 of the frequent target sequences and 11 of the infrequent targets. Also, the differences between the two target groups' scores for participant 8, 20, 27 (28) are close to significant (see Table 4).

Table 4: Participants' scores and P-values

	Target sequences		
Participants	Frequent targets[1]	Infrequent targets[2]	P-value*
1.	20	10	0.1758
2.	17	3	**0.0077**
3.	23	7	**0.0139**
4.	21	6	**0.0180**
5.	17	2	**0.0025**
6.	23	10	0.0793
7.	23	5	**0.0046**
8.	25	11	0.0586
9.	20	6	**0.0190**
10.	21	8	**0.0430**
11.	14	3	**0.0246**
12.	15	1	**0.0012**
13.	10	3	0.1222
14.	25	7	**0.0080**
15.	20	3	**0.0022**
16.	21	7	**0.0358**
17.	15	7	0.1416
18.	20	9	0.1101
19.	20	11	0.1874
20.	25	10	0.0516
21.	24	9	**0.0309**
22.	22	5	**0.0050**
23.	20	4	**0.0061**
24.	11	1	**0.0090**
25.	18	3	**0.0042**
26.	20	11	0.1874
27.	25	10	0.0516
28.	20	8	0.0652
29.	20	6	**0.0190**
30.	23	10	0.0793
Totally	598	196	< 0,0001

[1] Max = 29 [2] Max = 30 * Significant difference in bold

The participants' performances are illustrated in Figure 7. Each participant is represented with a number from 1-30 on the x-axis. Participant number 1 has correctly reproduced 20 frequent target sequences and 10 infrequent targets. The results show the tendency for the participants' score in each frequency group to correlate with each other. A relatively within-group high score for the frequent targets indicates a relatively, within-group, high score for the infrequent targets as well, while a relatively, within-group, low score for the frequent targets indicates a relatively, within-group, low score for the infrequent targets.

The participants' scores for the frequent target sequences are marked in red and the scores for the infrequent targets are marked in yellow. Association lines are added to illustrate the correlating within-group relative score for the two frequency groups for each participant.

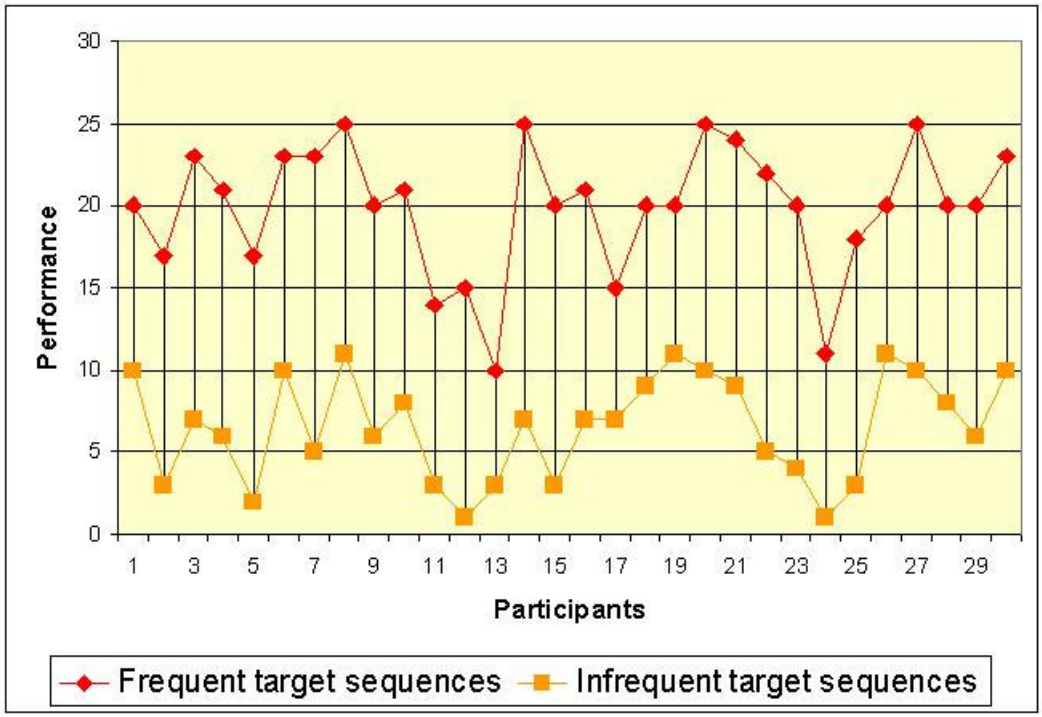


Figure 7 Participants' scores

Target sequences' scores

The scores for each of the target sequences are shown in Figure 8 and Figure 9 below. Maximum score was 30 correct reproductions for the 29 frequent target sequences and the 30 infrequent target sequences. The mean score for the frequent target sequences is 20.6 and 6.5 for the infrequent targets. Here the variation is greater, with a standard deviation of 8.3 for the frequent targets, and 5.5 for the infrequent targets. Nonetheless, there is a clear tendency

for the frequent target sequences to be correctly perceived more often (illustrated in Figure 8) than the infrequent target sequences (illustrated in Figure 9).

In Figure 8 and Figure 9 below, each of the target sequences are represented at the x-axis and their scores at the y-axis. Six of the frequent target sequences are correctly reproduced by all the participants, while five of the sequences are correctly reproduced less than ten times. The majority of the frequent targets were reproduced correctly by more than 20 of the participants.

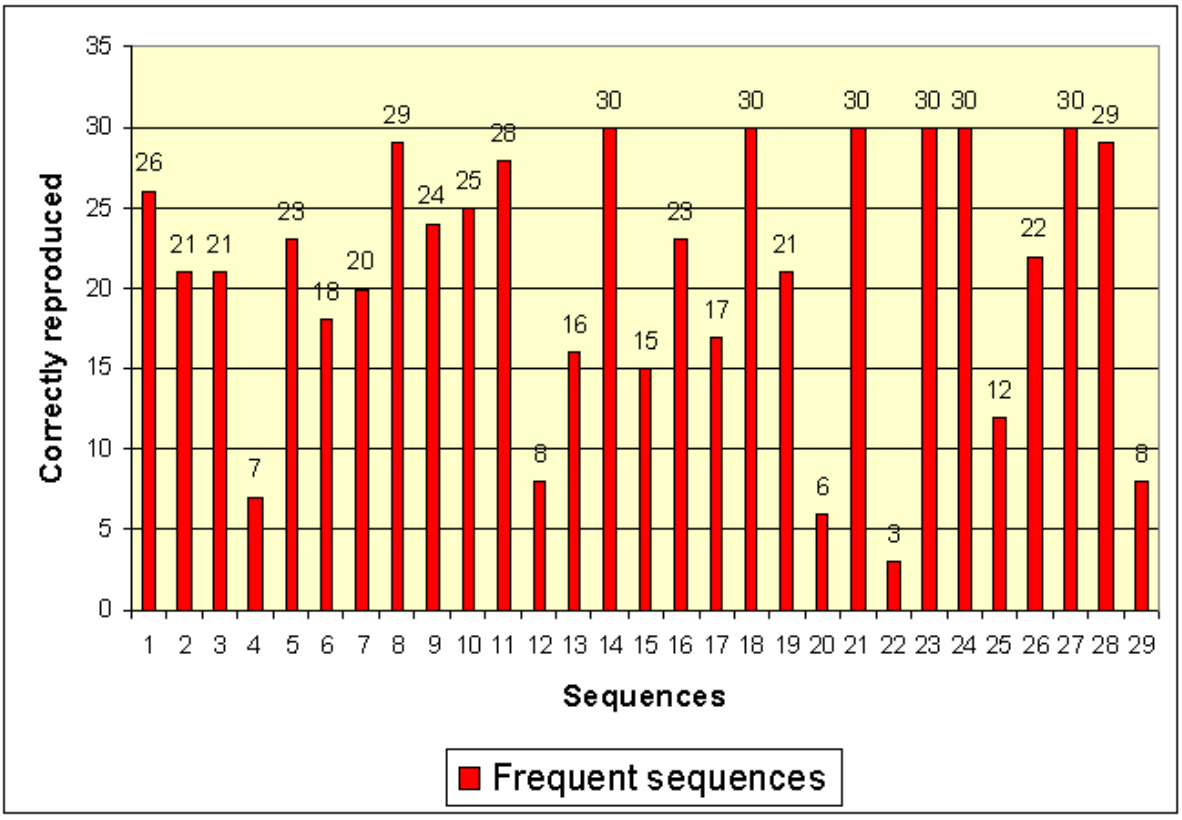


Figure 8 Frequent target sequences' scores

One of the infrequent target sequences stands out from the others. Sequence number 36, *Det er så jeg ikke* has been reproduced correctly by as many as 22 of the 30 participants; however, only six of the 30 infrequent target sequences have been correctly reproduced by more than ten of the participants.

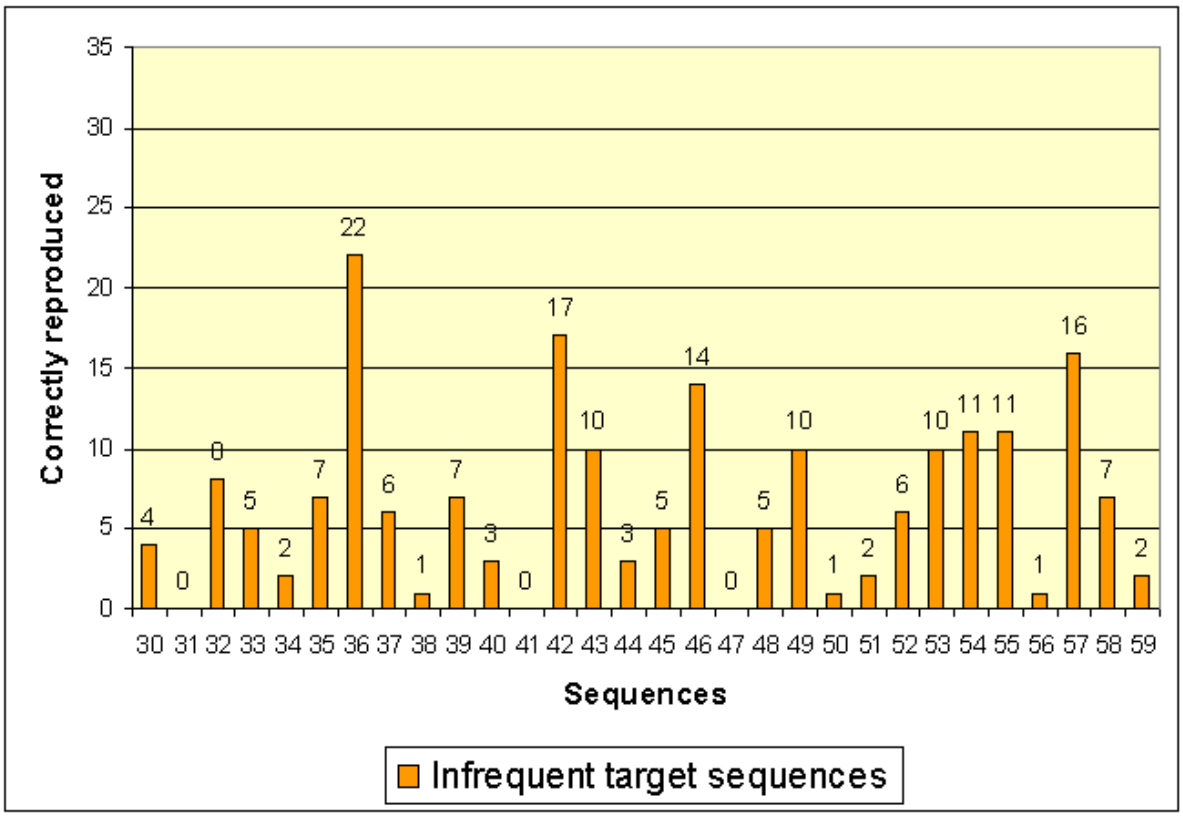


Figure 9 Infrequent target sequences' scores

4.4.2 Qualitative Results

In this section I present the qualitative properties of the responses from 27 recordings. Three of the participants were not recorded because of technical problems. Therefore the incorrect responses from these three participants are not included in this material.

The responses from all the participants have been orthographically transcribed and divided into six categories: 1) Correctly reproduced sequences, 2) Partly correctly reproduced sequences with one or more additional elements, 3) Partly correctly reproduced sequences with one or more missing elements, 4) Partly correctly reproduced sequences with one or more substituted elements, 5) Incorrectly reproduced sequences, and 6) No response. The categories are not mutually exclusive; some sequences miss elements from the stimulus and

at the same time include other elements not present in the stimulus. Still, the responses are divided in a fairly unproblematic manner, each into one primary category. The responses in categories 2–5 are given in its total in appendix III, while some examples are given below in the text.[28] I will also look at phonological and formal similarity and finally, participants' strategies. The partly correct and the incorrect responses are displayed in appendix III.

Correctly reproduced sequences

The participants tend to copy the intonation curve from the stimuli for the correctly reproduced frequent target sequences, and the sequences are articulated in a secure manner as intonation units. This is not the case for the correctly reproduced infrequent target sequences, which tend to be articulated with more hesitation, with an interrogative intonation, or with a following comment, like *maybe, I am not sure, something like that.*

Partly correctly reproduced sequences with one or more additional elements

The partly correct response with an additional element consists of responses to 6 (out of 29) frequent and 10 (out of 30) infrequent target sequences. The responses in this category include one additional element, placed in front of, in the middle of, or at the end of the stimulus target sequence. The additional element does not significantly change the meaning of the original sequence in any of the responses. In the frequent target group, the response is typically another frequent multi-word sequence, for example:

Sq. nr.	Target sequence		Response sequence
1.	Ja men det er jo	→	Ja men det er <u>det</u> jo
3.	Det er det som er	→	Det er <u>jo</u> det som er
6.	Det er i hvert fall	→	Det er <u>det</u> i hvert fall
14.	Det er klart det er	→	Det er klart <u>at</u> det er

Some of the frequent target sequences are clearly part of larger sequences in use, evident in the response:

[28] Due to limited space, I refer to appendix III for the English translations of the responses.

Sq. nr.	Target sequence		Response sequence
7.	Jeg tror ikke det er	→	Jeg tror ikke det er det/noe
19.	Jeg er veldig glad i	→	Jeg er veldig glad i deg/å/mat/det

The responses for the infrequent target sequences in this category mainly include one additional element that makes the sequence somewhat more acceptable, yet not necessarily more frequent:

Sq. nr.	Target sequence		Response sequence
33.	Det er det å da	→	Det er jo det å da
35.	På en måte er sånn	→	På en måte er det sånn
39.	Er ikke noe på det	→	Det er ikke noe på det

Two of the responses in this group were sequences of relatively high frequency on the Internet, approximately 12,300 occurrences for response nr. 50, and 23,100 occurrences for response nr. 54:[29]

Sq. nr.	Target sequence		Response sequence
50.	Og så er så mye	→	Og så er det så mye
54.	I hvert fall er sånn	→	I hvert fall er det sånn

Partly correctly reproduced sequences with one or more missing elements

The partly correctly reproduced sequences missing one or more elements from the input sequences consist of response to 14 frequent target sequences and 24 infrequent target sequences. Some of the elements that are missing in the response are in fact unaccentuated and nearly merged into the elements they precede or follow in the stimuli, for example:

Sq. nr.	Target sequence		Response sequence
4.	Nei jeg vet ikke jeg	→	Nei jeg vet ikke
”	”	→	Jeg vet ikke jeg
11.	Det er jo ikke det	→	Er jo ikke det
40.	Jeg tror det jeg var	→	Jeg tror det var
50.	Og så er så mye	→	Så er så mye

The response in this group is sub-classified into three groups, 1) missing one element, 2) 2- or 3-grams, and 3) fragments consisting of one word, or non-continuous words from the stimulus.

A tendency is that the participants reproduce the 2 and 3-grams (see Section 4.3.2) that are part of the larger sequence. These two- and three-word sequences are located at the beginning or at the end of the target sequences, and correspond to the high-frequent 2- and 3-grams extracted from the NoTa-Oslo corpus, and which were used to construct the infrequent four and five word target sequences. Approximately 1/3 of the responses in this group consist of only a 2- or a 3-gram; however, the overall number of responses containing a 2- or 3-gram from the stimuli target sequence is even higher, as we shall see below.

Partly correctly reproduced sequences with one or more substituted elements

The responses in this category range from sequences with two elements preserved from the stimuli and all other elements substituted with elements not present in the stimuli, to sequences with only one substituted element. This category of responses is by far the largest category, represented by response to 19 (out of 29) of the frequent target sequences and to all of the 30 infrequent target sequences. Below are two of the responses to frequent targets, the first includes two preserved elements, and the second includes only one substituted element:

Sq. nr.	Target sequence		Response sequence
5.	Det er i hvert fall	→	Jeg vil <u>hvert fall</u>
8.	Det er ikke så veldig	→	<u>Jeg</u> er ikke så veldig x 2[30]

[29]Approximately the number of occurrences in a search on <www.google.com>

[30] The number indicates how many times the specific sequence has been produced.

Typical for this category, and especially for the responses to the infrequent target sequences, is that the responses consist of preserved 2- or 3-grams, and some 4-grams from the stimuli:

Sq. nr.	Target sequence		Response sequence
10.	Og så er det jo	→	Nå er det jo
7.	Et eller annet sånt noe	→	Et eller annet så… sånn
26.	Det er ikke noe sånn	→	Det er ikke noe tvang/sak/sannhet/sant
30.	Det er sånn jeg syns	→	Det er sånn jeg ser
”	”	→	Det er sånn jeg har skrevet
31.	Og det var hvis du	→	Og det var dumt du
31.	Men det var jeg bare	→	Men det var jo en
34.	At det er at jeg	→	Han ser at jeg

Incorrectly reproduced sequences

The incorrectly reproduced sequences consist of responses that have none, or only one preserved element from the stimulus. When comparing the responses from the two frequency groups, the number of responses in this category is generally larger for the frequent targets than for the infrequent targets. The incorrect response, especially the response to infrequent target sequences, is mostly containing only one preserved element from the stimuli:

Sq. nr.	Target sequence		Response sequence
1.	Ja men det er jo	→	Man lærer jo
3.	Det er det som er	→	Jeg tror ikke det
4.	Nei jeg vet ikke jeg	→	Det er jo ikke
38.	Nei det var på det	→	Er det vannet x 3
”	”	→	Er det bable
47.	Ja og så er ikke	→	Ja sa hauken

In two cases, the responses might be said to resemble the target sequence semantically. This is a somewhat loose assumption, however the words *lyst* (‘delight’/’joy’) and *liker* (‘like’/‘enjoy’) both share the meaning of pleasure, and can for these specific sequences both be translated with the English word ‘like’:

Sq. nr.	Target sequence		Response sequence
9.	Jeg har lyst til å	→	Jeg liker det/deg

The sequences that have no preserved elements are in some cases guesses which to some extent resemble the stimulus' phonological profile:

Sq. nr.	Target sequence		Response sequence
12.	I og med at jeg	→	Mediet
20.	Holdt jeg på å si	→	Hotellet blir
”	”	→	Hotell og bil
”	”	→	Fra samme by
33.	Det er det å da	→	Jeg blir jo glad
45.	At det var hvis du	→	Kan se vanskelig ut

Participants' strategies

The responses are, to a large extent, mirroring the participants' response strategies. The participants can roughly be divided into two types, the "guessers" and the "passers". The guessers generally do not censor the perceived word or sequence, even if they are uncertain if it is correct – or even if they are certain that it indeed must be wrong. The passers, on the other hand, refuse to respond if they are not confident that the response is correct, or close to correct. Some of the participants responded in one- or two-word-manner, almost consistently, in the cases where they probably considered the perceived sequence to be incorrect. This may have impacted the results for the infrequent sequences, as the sequences may have been heard, but not reproduced. However, this does not change the fact that the frequent sequences were reproduced in a secure manner, and I thus consider this to be a qualitative difference between the two frequency sets.

4.4.3 The Results in Light of the Formal Dual-Mechanism Model

The prediction regarding the perception and reproduction of fully compositional (i.e. regular) multi-word sequences, according to the formal dual-mechanism model, is that the recurrent target sequences in the material should show no frequency effects. As long as the sequences' composite parts are controlled for frequency, the assumption is that there should be neither significant quantitative nor qualitative differences in the perception and reproduction of the different groups of target sequences.

With regard to the quantitative results of the experiment, the results go counter to the predictions given by the dual-mechanism model. The highly significant difference in the

number of correctly reproduced target sequences in the two groups of target multi-word sequences may indicate that the two groups have fundamentally different properties, leading to differences in processing.

A comparison between the targets and their responses also indicates a difference between the two target groups. While the responses for the frequent target sequences resemble the targets in fluency and intonation curves, the responses for the infrequent targets are more hesitant. However, since I have not carried out an in-depth examination of the phonological or intonational properties of the material and data, conclusions based on this data are less certain and must be treated with caution.

In sum, the results suggest that the formal dual-mechanism model cannot predict the right outcome for the experiment on perception and reproduction of recurrent multi-word sequences. The model cannot explain the fact that the fully regular compositional multi-word sequences are processed differently than infrequent and hence, supposedly non-storage sequences.

4.4.4 The results in light of the Usage-Based Associative Single-Mechanism model

The results of the psycholinguistic experiment are in accordance with the predictions given by the usage-based model: Frequency of use is assumed to affect multi-word sequences despite the lack of other properties associated with lexicalization. All things being equal except for frequency, a usage-based model predicts that the recurrent sequences will behave differently than the infrequent sequences, because of differences in degree of entrenchment of the sequences in the different frequency groups.

From a psycholinguistic perspective, the ART model (described in Section 3.4.1) hypothesize that correct perception of distorted speech signals to a large extent depends on a strong top-down support. The fact that the frequent targets are correctly reproduced a significantly higher amount of times, may therefore indicate that frequency of use leads to a strong mental representation for the entire sequence as a unit, enhancing processing.

The qualitative results show that the frequent target sequences are typically reproduced correctly, or partly correct with an additional element. Typical for the infrequent target

sequences is that they are partly correctly reproduced, conserving 2- or 3-grams from the stimuli. The fact that the largest group of partly correctly reproduced targets overall are sequences with 2- or 3-grams preserved from the targets, agrees with the usage-based account: these 2- and 3-grams are extremely frequent in use, which makes them easy to perceive compared to both the infrequent targets, and the frequent target sequences, which after all are relatively less frequent compared to the 2- and 3-grams.

5 Discussion

In this chapter, I relate the results of the psycholinguistic experiment to the research questions stated in Section 1.2 above, and evaluate the methodology and theoretical framework used in the present study. In Chapter 2, I presented and reviewed a study by Schmitt et al. (2004), which had the same starting point and focus as this study, and which thus makes up a natural comparison. The Schmitt et al.'s study came to contrasting results, a fact that calls for a discussion of the different methodological choices in these studies. The last part of the chapter includes a more general theoretical discussion of what the relation may be between frequency data from corpus and mental grammars.

5.1 The Findings Related to Research Questions

The working hypothesis of this study is that the recurrent sequences observable in language use are conventional units, represented as entrenched activation patterns in the language users' minds. From a usage-based linguistic view, the basic assumptions are that the repetition of linguistic structures leads to strengthened mental representations, and that conventional language enhances language comprehension and production. The frequent use of linguistic elements in the speech community in general is thus assumed to indicate high probability of mental storage.

5.1.1 Frequency of Use and Mental Representations

The psycholinguistic experiment was designed to isolate the variable frequency to investigate whether the frequently used, fully compositional multi-word sequences are represented as entrenched activation patterns in language users' minds. The results of the experiment (cf. Section 4.4) show that the two groups of target sequences are perceived and reproduced differently. Provided that the variable which was intended to be tested actually is the decisive factor, the results give a strong indication that frequency of use affects the mental representation of language units beyond the word, even for fully compositional multi-word sequences.

The results show that the frequent target sequences are correctly reproduced more than three times as often as the infrequent targets. The explanation for this is, according to usage-based theories and exemplar models of language storage and processing, that the recurrent multi-word sequences are represented as entrenched activation patterns in long term memory for language users that hear, speak and think these sequences frequently. Depending on the language users' unique experience with the target sequences, the individual sequences will be more or less entrenched. Some sequences will be strongly represented, and therefore more easily activated and retrieved than sequences which have a weaker representation or no mental representation at all. The difference in reproduction between the two frequency target groups is thus assumed to reflect a difference in mental representation: The infrequent target sequences are in general not perceived and reproduced by the participants because they have not heard or used these sequences before, or the sequences have a weak mental representation because of extremely low frequency of use for the individual in particular, and/or in the speech community in general.

The assumption that the frequent target sequences have a strong mental representation is also supported by the qualitative data. The reproductions of the frequent targets match the targets' intonation curve, and are generally articulated in a fluent manner. In contrast, the test participants' reproductions of the infrequent target sequences are characterized by hesitation and non-fluency. A possible explanation for the qualitatively different properties between the frequency groups is that the frequent target sequences are automatized units. Automatized complex units are typically uttered fluently and with preserved intonation profiles (cf. Section 4.2.2), and are assumed to constitute units of storage (cf. Section 3.3.2).

Other plausible explanations for the results of the experiment must of course be considered. It may be that variables other than frequency cause the quantitative difference in reproduction of the two target groups. One suggestion is that the material used as targets in the experiment may have properties other than frequency that affect the perception of the sequences. Regarding the recurrent sequences extracted from the NoTa-Oslo corpus of spoken language, they are fully compositional; however, they may still be assigned special pragmatic functions in the language. This is undoubtedly the case for target sequence nr. 20, *holdt jeg på å si* ('I almost said'), which follows an utterance that is indeed uttered. The sequence expresses the speaker's doubt towards the appropriateness or the relevance of the preceding utterance, and is not usually used in its literal meaning. Within the research field,

this type of pragmatically loaded sequences is regarded as formulaic. They are salient, and tied to more or less standardized communication situations (see Section 2.3.1).

Pragmatic function is a less tangible criterion; at least if the assessment is done without a thorough semantic analysis. For the purpose of the present thesis a shallow evaluation of the frequent target sequences, based on native speaker intuition, suggests that they often have specific functions in texts, i.e. to bind elements of a text together. Most of the sequences are used to introduce new facts, attitudes or beliefs, for example: *og så er det jo* ('And than there is'), *jeg har lyst til å* ('I would like to'); however they cannot be said to express any self-contained pragmatic meaning, exceeding the meaning of the sum of its parts. The sequences are not necessarily carrying the discourse functions they have in discourse, a fact suggesting that this discourse functional aspect is not an intrinsic property of these sequences (Tremblay, Derwing, Libben, and Westbury, 2008). The sequence *det er det som er* ('that's how it is') is a somewhat more self-contained pragmatic unit. Still, the pragmatic meaning is not as clear as for sequence nr. 20, mentioned above. Three of the 29 frequent target sequences are probable candidates for pragmatic units; Sequence nr 20: *holdt jeg på å si* ('I almost said'), nr. 2: *for å si det sånn* ('so to speak') and nr. 7: *et eller annet sånt noe* ('something like that') are represented in the NoTa-Oslo several times in isolation, expressing a self-contained pragmatic meaning. However, apart from these three examples, the recurrent multi-word sequences in the experiment are rather non-salient, both semantically and pragmatically. To paraphrase a given sequence is difficult without exemplifying a situational context. For that reason, even though the recurrent multi-word sequences may be functional units in conversation, they are, for the most part, not typical pragmatic units. This speaks in favour of the assumption that the sequences have strong mental representations, which makes them easier to perceive and reproduce than the infrequent target sequences. This fact is suggested to be a consequence of frequent use alone, because the sequences do not have specific pragmatic meanings.

Another fact supporting these assumptions is that an overwhelmingly high amount of the correct responses correlate with the frequent 2 and 3 grams represented in the NoTa-Oslo corpus. In the psycholinguistic experiment, these sequences are elements within the frequent target sequences as well as within the infrequent target sequences, for example *det er jo* (1977 occurrences in NoTa-Oslo) ('well, it is…'), *ja det er* (1341 occurrences) ('yes, it is…), *ja men det* (512 occurrences) (yes, but it…). A majority of these sequences are not pragmatic

self-contained sequences; they are just extremely frequent in language use. This fact may be seen as supporting the hypothesis that frequency of use leads to the entrenchment of sequences beyond the word, and again, even for sequences that do not have any specific pragmatic meaning attached to them.

A possible objection to the material in the present experiment is that the perception and reproduction of the frequent target sequences indeed reflect ordinary rule based computation, while it is the infrequent target sequences that cause perception problems for the test participants. Is it possible that the infrequent targets are infrequent in use because they are less grammatical and hence, unacceptable sequences which make them hard to perceive? The answer to this question is both yes and no. The fact that the sequences seem to be less acceptable is a consequence of a low degree of conventionality. They seem somewhat odd, and seen out of normal context, the sequences are rated as less acceptable. However, it may be argued that this has nothing to do with the sequences' possibility or impossibility, as Sampson (2007: 5) states it:

> *The community is simply progressively discovering more and more ways to achieve rhetorical goals by putting words together, and although this is a process that unrolls through time so that not all possibilities are well-established at any given date, there is no reason to think that any particular sequences of words are definitely "out of bounds" at a given date – perhaps in fact it will not happen to be for several decades before someone first finds a use for sequence* X*, but it could be today.*

Most of, although not all, the infrequent target sequences are used by some language users, they are intelligible and, most importantly, they are rated by other language users as acceptable sequences in specific contexts (cf. Section 4.3.2). Even though these sequences seem to be less preferable ways to express meaning contents, nothing prevents them from becoming conventional in the future. The reason that this may indeed happen, is that speakers tend to echo features of the speech of their interlocutor(s). This tendency is termed accommodation (Field, 2004), and includes the echoing of words, sequences and syntactic patterns, as well as speech style and accent. Also, according to usage-based theories, any previously activated linguistic elements will more probably be repeated – which contributes to strengthen these elements' mental representations (cf. the feedback loop in Section 3.2, Figure 3).

Even though it is assessed here that high frequency leads to entrenchment of multi-word sequences, it is important to stress that high frequency is not the only decisive criterion. Low frequency units may also very well be storage and processing units, since other properties of the sequences, besides recurrence, also affect the linguistic units' representation in mind (cf. Section 3.3.1). A one-sided focus on frequency in the search for complex storage units will probably only capture part of the picture. However, there is no reason to go to the other extreme either. Even though high frequency is not a necessary criterion for a sequence to possess a strong unitary mental representation, one cannot assert that frequency of use is irrelevant. Rather, salience and frequency are two separate routs to entrenchment. While both types of sequences may have strong mental representation, the salient sequences, or units, do not need to be especially frequently used – nor are the frequent units necessarily especially salient. The point is that frequency of use is reflected in the processing of sequences. This property of the sequences must be seen independent of other properties. The present study shows that frequency affects processing systematically and therefore there is reason to assume that the mental representations of linguistic structures, beyond single words, are strengthened as a consequence of frequent use. The next section questions which of the two competing mental models presented in Chapter 3 is compatible with these findings.

5.1.2 Dual or Single Mechanism Model?

Two competing models of language storage and processing were presented in Chapter 3. Different predictions regarding the perception and reproduction of multi-word sequences in two different frequency ranges were deduced, based on the models' structural and processing principles. In this section, I consider whether the results from the present psycholinguistic experiment are in line with the predictions deduced from the dual-mechanism model, or if the associative, single mechanism model better predicts the outcome of the experiment.

The generative dual-mechanism model asserts that because the multi-word sequences from both frequency groups are fully compositional, all the target sequences will be processed by the same rule based computational system. Therefore, the assumption is that the sequences' frequency of use is irrelevant regarding processing, as long as both groups of target sequences are adjusted for word frequency. Thus, any significant difference in processing properties between the two groups of target sequences is not anticipated. The results do not,

however, support this prediction, since frequency effects are indeed found for the fully compositional multi-word sequences. This fact indicates that multi-word sequences need not be irregular to be entrenched linguistic units. Models of language storage and processing, claiming to be psycholinguistically valid models of linguistic representation, need to take into account the fact that frequency of use affects linguistic structures of all types and sizes. Psycholinguistically valid models thus require redundant storage, which is not compatible with a principled division between a rule based and a memory based computational system, based on the economy principle. "First, you must account for the data. If you have two hypotheses which account for the data equally well, then you may consider economy" (Lamb 1999: 128). The results of the present experiment suggest that the dual mechanism model fails to predict the processing properties of the two target groups. Moreover, the results also suggest that the model fails to account for the data, because the model assumes that frequency of use for fully compositional units beyond the word is an insignificant fact, not affecting the mental representations for this kind of linguistic units. The generative dual-mechanism model's lack of explanatory power is contrasted by the usage-based, single-mechanism model, which is able to predict the differences in processing of multi-word sequences based on differences in frequency, and which naturally includes and expects the phenomenon of recurrent sequences.

I propose that one problem with defining the phenomena of formulaicity is partly that grammatical theories (in general) have a strong focus on the production side of the linguistic system (see Lamb, 1999). Often, assumptions of what is supposed to be stored in the mental grammar are based on what language users need to know to *produce* novel utterances. This is of course an important aspect of the linguistic system; however, the ability to *comprehend* language is undoubtedly prior to production, a fact which is evident in language learning. From a usage-based viewpoint, the receptive system is the totality of nodes and connections which is present in language users' language memory system. This is what makes us able to comprehend even low frequency and sparkling new sequences of words, while the ability to produce the same sequences is more restricted because this active process is more demanding, and is also more dependent of strong mental representations. The point is that the receptive system does not favour the smallest units; to comprehend complex sequences of words, or even larger units at discourse level, the storage of larger activation patterns facilitates the perception process. It is this receptive system that should be investigated in

relation to the question of storage or non-storage of formulaic sequences, and it should also be used on a more general basis for the description of the mental lexicon.

For the purpose of linguistic description, there is certainly a need to divide language structures into different classes. This division should not, however, be taken as reflecting psychologically real structures. Proponents of dual-mechanism models argue that a difference in processing between complex lexical units and compositional sequences is evidence for a dual processing system; however one should be cautious to assume that if different language structures behave differently, there is reason to postulate different mental components to account for these differences. In other words; the fact that linguistic structures are differently processed does not entail different processing systems. A more plausible explanation may be that the difference in processing reflects different properties of the linguistic structure, while the processing system is aimed at handling tasks of both declarative and procedural nature. Supporting this assumption is the fact that even low frequency idiomatic sequences may be productive, cf. the examples *I will eat my hat* and *Is the Pope Catholic?* in Section 2.2, which are idiomatic sequences, but which still are subjects of analogy. The ability to make extensions of complex lexical units is evidence that the lexicon is not a static storage of pure declarative knowledge, but a dynamic system with constructions of varying size and complexity, and which forms patterns for the recurrent, the extended or the creative use of language.

Within an associative, single mechanism model, the recurrent sequences are represented as strong activation patterns which are easily activated, both in perception and production, and which also are productive. Variations over the same patterns are evident in corpora of language use (cf. the associative network illustrated in Figure 5), and indicate that the entrenched activation patterns are part of the language memory system as resources for both speakers and hearers. It may be assumed then, that the repeated patterns of language use are not arbitrary, but reflect a complex system of linguistic structures which is the result of prior usage (the receptive system), and which is the subject of usage in the present time, and will be in future times (the productive system). The usage-based, single-mechanism model is accordingly able to account for both prefabricated as well as creative language use within the same computational system.

5.2 The Present Study Compared to Earlier Studies

The studies by Schmitt et al. (2004) and Vogel Sosa and MacFarlane (2002), outlined in Chapter 2, are both studies of recurrent sequences with a psycholinguistic approach, aimed at testing their mental representations, partly in relation to frequency of use.[31] The studies came to contradictory conclusions regarding the relation between frequency and mental representation. In this section, I compare these studies to the present study, and evaluations of the different methods used are given.

In their study "Are corpus-derived recurrent clusters psycholinguistically valid?", Schmitt et al. (outlined in Chapter 2) incorporated 20 recurrent sequences of different sorts in a dictation task. Amongst these, several of the sequences were characterised as fully compositional. These sequences are frequently used in speech or writing; however, they lack any of the specific properties traditionally associated with formulaicity. This type of recurrent sequence is the focus in the present thesis, and the results for these sequences in the study by Schmitt et al. are thus particularly interesting for comparison. Schmitt et al. analyse their results in reference to the traditionally defined properties associated with formulaicity. The fact that the fully compositional recurrent multi-word sequences in their study are reproduced less frequently compared to sequences with other properties, besides recurrence, leads to a conclusion that frequency alone does not necessarily lead to entrenchment of fully compositional multi-word sequences. This conclusion is in accordance with the research field's prevailing assumption that frequency is not a decisive factor (see Chapter 2), which thus seems to be further supported through this study.

The lack of correlating results for these studies is not rooted in a difference of the materials that have been used, but in different methodological choices. It is difficult to design a study that exclusively tests the intended variables, so most, if not all studies have their weak points. The way I see it, Schmitt et al. are testing other variables than they intended. Their methodological choices produce data that may say something interesting about differences in salience between the different sequences; however, I question the relevance of this study in relation to the sequences' storage properties.

My first objection to Schmitt et al.'s study is their choice to use a narrative as background for the target sequences. In their study, they stated that "It was felt desirable to have the dictation bursts form a coherent text, rather than be series of unrelated bursts…" (Schmitt et al. 2004: 131). They give no explanation for this desire; however in Conklin and Schmitt (2007) they present the same argument, based on the fact that in actual language use, "formulaic sequences do not exist in isolation, but rather in discourse" (Conklin and Schmitt, 2007: 7). Therefore, they argue that the sequences should be inserted in an appropriate context to enhance perception of the sequences. This assumption is also held by Goldinger (1998: 268), who states that mental representations are not "perceptual analogues, totally defined by stimulus properties". The representations are rather "perceptual-cognitive" objects, jointly specified by perceptual forms and cognitive functions" (ibid.). This might explain why obviously known elements are not perceived if encountered in an unknown or unexpected context. An example is the problem of comprehending an utterance if it is spoken in a known, yet unexpected language. Despite these plausible reasons for presenting the target sequences in context, my objection to this specific methodological choice is that the context, which is a narrative, opens up for other variables that most likely affect the results. Schmitt et al. themselves point to this possibility in their discussion of the results for the clusters *from the point of view, in addition to the, aim of this study, in the number of,* and *as shown in figure* (cf. Section 2.3.3). These clusters were more difficult for the participants to recall. As a result, they found it probable that because the clusters seem to point to a formal and academic register, they make a breach with the more informal tone of the narrative. Since the narrative is in focus when the participants try to recall the dictation bursts, I find it likely that some of the clusters that breach with the register, or do not contribute to the semantic/pragmatic coherence of the story, are either omitted or repeated incorrectly. In other words, by inserting the sequences into "wrong" context, the context fails to prime the target sequence.

Another objection to Schmitt et al.'s study is one concerning function. Because the study aims at testing the participants' reproduction of sequences incorporated in a coherent text, they expect the participants to be able to recollect a specific sequence from memory that fills

[31] The study by Vogel Sosa and MacFarlane is examining *collocations*; however, collocations are also *recurrent sequences*. Collocations are just a specific type of recurrent sequences.

a specific function, or a specific semantic task in the text. The problem with this assumption is that a specific function does not necessarily invoke one and only one specific word sequence, but potentially a whole range of sequences. In my opinion, the common one sided focus on production is to blame (cf. Section 5.1.2). In order to investigate the inventory of multi-word sequences in mind, it may be more advantageous to use recognition rather than production as a basis for the investigations, which is exactly the methodology used in the present psycholinguistic experiment. The reason for choosing recognition over production is that we easily make the connection between a linguistic expression and its meaning. For entrenched sequences, this process is a more direct one than for non-stored sequences. However, the process of finding the right sequence to express one specific idea is a more deliberate process (Lamb, 1999: 132). When the sequences in addition do not contribute directly to the coherent story, which is the case for some of the sequences in Schmitt et al.'s study, there is reason to believe that the test subjects in the study have problems with retrieving even probably stored units.

I view their experiment as only partially successful. The recurrent clusters that all or most of the participants reproduce are asserted to have a probable status as memory units in mind; however, the recurrent clusters that are assumed to either fall into the class of less probable or not probable storage units, may still be likely storage units if tested in a refined experiment.

A study with correlating results compared to the present findings is Vogel Sosa and MacFarlane's study from 2002 of collocations including the word *of* (outlined in Chapter 2). Their hypothesis was that the reaction time to utterances containing collocations of high frequency should be *longer* because the sequences' compositionality is reduced. The reduction is a consequence of frequent use, both phonologically and semantically, causing the elements of the sequences to become semantically bleached, or less tangible. Therefore they expected that the elements *of* within the reduced sequences would be more difficult to spot. The results show that the mean reaction times to the highly frequent collocations are significantly higher than the reaction times to the less frequent collocations. They conclude that this indicates storage of the two-word collocations in that group, and that the significance can be attributed to frequency effects. Vogel Sosa and MacFarlane state that their results are consistent with usage-based models of language storage and processing, which is in line with the findings and the theoretical considerations in the present study.

5.3 Corpus Data and Mental Grammars

In this thesis I have argued that recurrent sequences extracted from an oral corpus are entrenched sequences in the language users' minds. Further, I have also stressed that each language user's knowledge of linguistic units and grammar is highly individual (see Chapter 3). These assertions may seem self-contradictory, and it is therefore natural to ask if we can infer from language use in a community to the individual language user's mental linguistic structures.

The link between language use as represented in corpora, and the individual language user's mental grammars is obviously not a direct one. Still, by assuming a usage-based theoretical approach which sees the mental representation of language as a result of language usage, and language use as a means for communication, there is good reason for investigating the potential advantages of a corpus-based approach to mental grammars. Because every person's linguistic experience differs from other persons' experiences, it follows that no grammatical systems – which are assumed here to be the results of linguistic experience, are identical. Nonetheless, large amounts of linguistic experience are shared by groups of language users, and therefore, the conventional aspects of language are assumed to be reflected in language production and in corpora of language use, and also in the minds of the language users, which are representative of the speech society in question. While specific units may be parts of the lexicons for only a small group of language users, i.e. words and phrases like *creolization, agrammatism, Zipf's law* and *top-down processing*, the recurrent sequences found in corpora like NoTa-Oslo are probably deeply entrenched activation patters in most, if not all, speakers representative of a particular speech community, despite the fact that these sequences are fully compositional, literal and analyzable.

Usage-based models are concerned with patterns of language use and whether these patterns are common or rare, rather than asking if a specific pattern confirms to a predefined correct grammar. Therefore, "adequate investigations of language use must be empirical, analyzing the functions and distribution of linguistic features in natural discourse contexts" (Biber, 2000: 287). Corpora are well suited for investigating actual patterns of language use, and especially to establish surveys over conventional multi-word sequences in a given language variety. Other methods must of course be used in addition to capture the entire range of formulaic sequences; however, the conventional patterns which are available from corpus

based analyses may have a specific advantage in being possible, and also probable, elements in a representative associative network of complex linguistic elements.

6 Summary and Conclusions

The main goals of this thesis have been to extend the category of formulaic sequences to include also the conventional recurrent and fully compositional multi-word sequences evident in language use, and to evaluate two competing models' ability to predict and explain the processing properties of this kind of recurrent multi-word sequences compared to infrequent sequences. I will conclude this thesis by summarizing the main findings related to the thesis' goals. Also, I will give some suggestions for further research on the relation between frequency, corpus and mental representations to follow up on the findings of the present study.

6.1 Implications for the Research Field

In Chapter 2, I gave a review of the research field on formulaicity, and pointed at a generally recognized problem of categorizing and delimiting the phenomenon. The psycholinguistically based definition of formulaic sequences by Wray (2002) has been used as the reference point for examining whether the recurrent sequences extracted from the Norwegian oral corpus NoTa-Oslo (cf. Section 4.2) are probable storage units, and thus also formulaic sequences by this definition. The present findings show that sequences which lack the properties traditionally defining formulaicity, and which therefore fail to be categorized as formulaic sequences, still are probable storage units. The prevailing assumption within the research field, that there exists a principled division between stored multi-word sequences and non-storage sequences based on criterial features traditionally identified with formulaicity, is therefore suggested to be erroneous.

Within the research field, there is general agreement that because the recurrent and fully compositional multi-word sequences lack properties traditionally identified with formulaicity, these sequences cannot be storage and processing units. However, the mere fact that these sequences are not in possession of specific properties does not denote that they are not stored and processed as entrenched activation patterns. It only means that the sequences do not possess these other specific properties. The question of storage and processing is better investigated from a psycholinguistic perspective, and the present findings indicate that

the research field should include also the recurrent and fully compositional multi-word sequences to the class of formulaic sequences – as long as formulaic sequences are defined as storage and processing units (cf. Chapter 2). In order to be able to define the criteria that distinguish formulaic sequences from non-formulaic sequences, it is important to operate with all types of storage and processing units and not only the units located towards the more idiomatic end of the scale. This implies that the distinction between what may be and what may not be units of storage cannot be a principled one, but rather is a question of probability. This is not compatible with the classical, Aristotelian way of categorizing, based on necessary and sufficient criteria (cf. Section 2.3) which is the categorizing principle applied within the research field of formulaicity. I propose that an exemplar-based network of more or less entrenched activation patterns, where the most entrenched units function as central members (cf. prototype theory, described in Section 2.3.1), better captures the storage and processing properties of complex units, and that such a network is a better starting point to establish the criteria identifying formulaicity.

Much research on formulaicity has a strong focus on the challenges which learners of languages meet when they encounter complex and non-transparent language, and the mastering of formulaic language is indeed an important aspect of mastering a language. However, competent language users also need to know the conventional units. Conventional language contributes to ease and promote communication, as language users guide each other through language by means of familiar language paths. The research field should therefore benefit from including recurrent multi-word sequences, which are linguistic challenges that become resources when they are first mastered.

6.2 Theoretical Implications

In this thesis, I have argued that the fact that recurrent, but fully compositional sequences show processing benefits compared to infrequent sequences indicates that the individual language user's grammars are, to a considerable extent, formed by the language users' linguistic surroundings and use. I therefore propose that language users' mental grammars reflect the individual's linguistic experience such that low frequency (morphemes), words

and sequences of words, dependent upon the linguistic structure's salience (cf. Chapter 3), are represented as weak activation patterns.[32] Items within the medium high frequency range have a yet stronger representation, while high-frequency units have deeply entrenched representations. Perception of low frequency items is possibly only accessible through recognition through a strong, clear and/or correct bottom-up representation, while high-frequency units have a strong top-down representation, and are thus less dependent on a strong bottom-up signal to be perceived, and are also strong candidates for further production. While other aspects than frequency are indeed relevant for, and forming the linguistic system, the fact that frequency affects the mental representations of complex units must be incorporated in a psycholinguistically valid model of grammar.

Based on my findings, I question the traditional principled distinction between open and bound constructions, because even idiomatic expressions are, to a certain extent, productive (cf. Section 2.2). Consequently, they must be the subject of analysis. This is not compatible with the traditional distinction between compositional and noncompositional sequences, reflecting separate computational systems. The present findings instead suggest that an associative, single-mechanism linguistic system is better suited to account for the fact that linguistic structures with different properties are differently perceived and processed, without postulating separate storage and processing mechanisms.

This thesis is ultimately about language users' knowledge of grammar and the way linguists model this knowledge within mentalistic approaches. If you advocate a mentalistic approach to grammar, you wish to posit grammars that are psychologically adequate, that is, all postulated structures, principles and processes are assumed to refer to psychological entities. The mentalistic models thus give testable predictions regarding processing properties of linguistic structures. The competing predictions deduced from an associative, single mechanism model and a dual mechanism model are the subjects of scrutiny in the present study.

Regarding the superiority of the usage-based, associative, single-mechanism model in the present study, it does of course not imply that this specific kind of model is the ultimate

[32] Low frequency units, which are also salient, are assumed to have a strong(er) representation.

psycholinguistically realistic model. Other models may equally well explain the results of the psycholinguistic experiment. What I have shown is that a dual-mechanism model is not capable of explaining the results of the present psycholinguistic experiment, a fact indicating that the model is not suited to explain a well known linguistic phenomenon. The model lacks explanatory power, and cannot be regarded as a realistic model of language storage and processing. I have shown that a psycholinguistically valid model needs to be usage-based and that a unitary system better explains and predicts the linguistic phenomenon in question. Recurrence is a natural part of the cognitive system, and naturally resulting from the usage-based linguistic system. In a natural way the usage-based model reflects language use – within the language society in general and in the individual language users.

6.3 Further Research

The psycholinguistic aspect of multi-word sequences is a relatively unexplored field since most research on frequency and lexical retrieval has focused on words. Possibilities for further research on the mental basis of multi-word sequences should thus be vast.

Because only sequences located at each end of the frequency scale have been included, the empirical study conducted in this thesis had limited scope. The results of the experiment show that the sequences are processed differently, a fact attributed to differences in frequency; nonetheless, to be sure that the alternative explanations for the test results (cf. Section 5.1.1) are rejected, a refined experiment including multi-word sequences in several frequency ranges should be conducted. According to the hypothesis stated in this thesis, the sequences' relative frequencies should be reflected in processing efforts.

While the present thesis has mainly focused on the quantitative differences in processing between the two frequency groups, the experiment generated lots of qualitative data which could be used as material for studies of processing. The present study does not explore in detail what the participants actually do when they produce incorrect responses: Do the incorrect responses reflect an underlying associated network of activation patterns?

Also, a closer look at the recurrent and fully compositional multi – word sequences in relation to aspects of productivity – within a usage-based theoretical frame, would contribute to a more comprehensive view of these types of sequences.

Bibliography

Barlow, S. and Kemmer, M. (eds.), 2000, *Usage Based Models of Language,* Stanford, California: CSLI Publications.

Biber, D., 2000, "Investigating Language Use through Corpus-Based Analysis of Association Patterns" in Barlow, S. and Kemmer, M. (eds.), *Usage Based Models of Language,* Stanford, California: CSLI Publications, 287–313.

BigBrother-korpuset, Tekstlaboratoriet, ILN, University of Oslo, [online], available at: <http://www.tekstlab.uio.no/nota/bigbrother/>

Bybee. J., 1985, *Morphology: a study of the relation between meaning and form,* Philadelphia: Benjamins.

Bybee, J., 1998, "The emergent lexicon", *Chicago Linguistic Society,* vol. 34: The Panels, 1998, 421-435

Bybee, J., 2001, *Phonology and language use*, Cambridge: Cambridge University Press.

Bybee, J., 2006, "From usage to grammar: the mind's response to repetition", *Language,* vol. 82(4), 711-733.

Bybee, J., 2007, *Frequency of use and the organization of language*, Oxford: Oxford University Press.

Bybee, J. and McClelland, J. L., 2005, "Alternatives to the combinatorial paradigm of linguistic theory based on domain general principles of human cognition", *The Linguistic Review,* vol. 22(2-4)*,* 381-410.

Carrol, D. W., 2004, *Psychology of Language,* 4. ed. Australia, Canada, Mexico, Singapore, Spain, United Kingdom, United States: Wadsworth/Thompson Learning.

Chomsky, N., 1972, *Language and mind*, New York: Harcourt Brace Jovanovich.

Christiansen, M. H. and Chater, N. (eds.), 2001, *Connectionist Psycholinguistics*, Westport, Connecticut and London: Ablex Publishing.

Conklin, K. and Schmitt, N., 2008, "Formulaic Sequences: Are They Processed More Quickly than Nonformulaic Language by Native and Nonnative Speakers?", *Applied Linguistics,* vol. 29(1), 72-89.

Coulmas, F., 1981, "Introduction: conversational routine", in Coulmas, F. (ed.), *Conversational routine: explorations in standardized communication situations and prepatterned speech,* The Hague: Mouton, 1-17.

Croft, W., 2001, *Radical Construction Grammar: Syntactic Theory in Typological Perspective,* Oxford: Oxford University Press.

Crystal, D., 1997, *A Dictionary of Linguistics and Phonetics,* 4. ed., Oxford and Malden, Massachusetts: Blackwell Publishers Ltd/Inc.

Dysthe, O. Hertzberg, F. and Hoel, T. L., 2000, *Skrive for å lære,* Oslo: Abstrakt Forlag.

Evans, V. and Green, M., 2006, *Cognitive Linguistics: An introduction,* Edinburgh: Edinburgh University Press.

Field, J., 2004, *Psycholinguistics: The Key Concepts,* London and New York: Routledge.

Giora, R. and Fein, O., 1999, "Irony: Context and salience", *Metaphor and Symbol*, vol.14(4), 241-257.

Givón, T., 1989, *Mind, code and context,* Hillsdale, NJ: Lawrence Erlbaum.

Goldberg, A., 1995, *Constructions: A construction Grammar Approach to Argument Structure*, Chicago: Chicago University Press.

Goldberg, A., 2006, *Constructions at Work,* Oxford: Oxford University Press.

Goldinger, S. D., 1998, "Echoes of Echoes? An Episodic Theory of Lexical Access", *Psychological Review,* vol. 105 (2), 251-279.

Goldinger, S. D. and Azuma, T., 2003, "Puzzle-solving science: The quixotic quest for units in speech perception", *Journal of Phonetics,* vol. 31, 305-320.

Goldinger, S. D. and Azuma, T., 2004, "Episodic memory reflected in printed word naming", *Psychonomic Bulletin & Review,* vol. 11, 716-722.

Grossberg, S., 1980, "How does a brain build a cognitive code?", *Psychological Review*, vol. 87, 1-51.

Howarth, P., 1998, "Phraseology and second language proficiency", *Applied Linguistics,* vol. 19(1), 24-44.

Hudson, J., 1998, *Perspectives on fixedness : applied and theoretical,* Lund: Lund University Press.

Jackendoff, R., 1995, "The boundaries of the lexicon", in Everaert, M., van der Linden, E., Schenk, A., and Schreuder, R. (eds.), *Idioms: Structural and Psychological Perspectives,* Hillsdale NJ: Earlbaum, 133-166.

Johannessen, J. B., 2003, "Innsamling av språklige data: Informanter, introspeksjon og korpus" in Johannessen, J. B., *På språkjakt,* Oslo: Unipub Forlag.

Lamb, S., 1999, *Pathways of the brain: The Neurocognitive Basis of Language,* Amsterdam and Philadelphia: John Benjamins Publishing Company.

Langacker, R. W., 1987, *Foundations of Cognitive Grammar, volume 1, Theoretical prerequisites,* vol. 1, Stanford, California: Stanford University Press.

Langacker, R. W., 1991, *Foundations of Cognitive Grammar, volume 2, Descriptive Application,* Stanford, California: Stanford University Press.

Langacker, R. W., 2000, "A Dynamic Usage-Based Model" in Barlow, S. and Kemmer, M. (eds.), *Usage Based Models of Language,* Stanford, California: CSLI Publications, 1-63.

Logan, G. D., 1988, "Toward an Instance Theory of Automatization" *Psychological Review,* vol. 95(4), 492-527, [online 2009.05.05], available at: <psych.wisc.edu/ugstudies/psych733/logan_1988.pdf>

MacWhinney, B., 2000, "Connectionism and Language Learning" in Barlow, S. and Kemmer, M. (eds.), 2000, *Usage Based Models of Language,* Stanford, California: CSLI Publications, 121-149.

Melcuk, I., 1998, "Collocations and lexical functions" in Cowie A. P. (ed.) *Phraseology: theory, analysis and applications,* Oxford: Clarendon Press, 23-53.

Merlo, P. and Stevenson, S. (eds.), 2002, *The Lexical Basis of Sentence Processing: Formal, computational and experimental issues*, Amsterdam and Philadelphia: John Benjamin's Publishing Company.

Moon, R., 1998, *Fixed expressions and idioms in English,* Oxford: Clarendon Press.

Nattinger, J. R. and DeCarrico, J. S., 1992, *Lexical phrases and language teaching,* Oxford: Oxford University Press.

Nordahl, H., 1994, *Retorikk: De viktigste retoriske figurer belyst ved eksempler fra riksmålets litteratur*, Oslo: Grøndahl Dreyer.

Norsk talespråkskorpus - Oslodelen, Tekstlaboratoriet, ILN, University of Oslo, [online], available at: <http://www.tekstlab.uio.no/nota/oslo/index.html>

Pawley, A. and Syder F.H., 1983, "Two puzzles for linguistic theory: Nativelike selection and nativelike fluency" in Richards, J.C. and Schmidt, R.W. (eds.), *Language and communication,* New York: Longman, 191–226.

Poulsen, S., 2005, "Collocations as a language resource: A functional and cognitive study in English phraseology", Doctoral dissertation, University of Southern Denmark, [online 2009.05.05], available at: <www.sdu.dk/~/media/98A7531AABC944388AE56628D9E4EBE9.ashx>

Read, J. Nation, P., 2004, "Measurement of formulaic sequences" in Schmitt, N. (ed.) *Formulaic Sequences: Acquisition, Processing and Use,* Amsterdam: John Benjamins, 23-36.

Sampson, G., 2007, "Grammar without grammaticality", *Corpus Linguistics and Linguistic Theory* 2007, vol. 3(1) [online 2009.05.05], available at: <http://www.reference-global.com/doi/abs/10.1515/CLLT.2007.001>

Schmitt, N. (ed.), 2004, *Formulaic Sequences: Acquisition, Processing and Use,* Amsterdam: John Benjamins.

Schmitt, N., and Carter, R., 2004, "Formulaic sequences in action: An introduction" in Schmitt, N. (ed.), *Formulaic Sequences: Acquisition, Processing and Use,* Amsterdam: John Benjamins, 1-22.

Schmitt, N., Grandage, S. and Adolphs, S., 2004, "Are corpus-derived recurrent clusters psycholinguistically valid?" in Schmitt, N. (ed.), *Formulaic Sequences: Acquisition, Processing and Use,* Amsterdam: John Benjamins, 127-152.

Schmitt, N. and Underwood, G., 2004, "Exploring the processing of formulaic sequences through a self-paced reading task" in Schmitt, N. (ed.), *Formulaic Sequences: Acquisition, Processing and Use,* Amsterdam: John Benjamins, 173-190.

Sinclair, J. McH., 1991, *Corpus, concordance, collocation,* Oxford: Oxford University Press.

Tomasello, M., 2003, *Constructing a Language: A Usage-Based Theory of Language Acquisition*, Cambridge, MA: Harvard University Press.

Tremblay, A., Derwing, B., Libben, G., and Westbury, C., 2008, "Processing Advantages of Lexical Bundles: Evidence from Self-Paced Reading Experiments, Word and Sentence Recall tasks, and Off-Line Semantic Ratings", manuscript submitted for publication, [online 2009.05.05], available at: <http://www.ualberta.ca/~antoinet/Processing%20advantages%20of%20LBs%20(19-01-2008).pdf>

Tremblay, A., Libben, G., Derwing, B., and Baayen, R. H., 2008, "Regular four-word sequences: An ERP study of the effects of structure, frequency, and probability on immediate free recall", manuscript submitted for publication, [online 2009.05.05], available at: <http://www.ualberta.ca/~antoinet/ERPDraft.pdf>

Tyler, A. and Evans, V., 2003, *The Semantics of English Prepositions: Spatial Scenes, Embodied Meaning and Cognition,* Cambridge: Cambridge University Press.

Ullman, M., 2001, "A neurocognitive perspective on language: the declarative/procedural model." *Nature Reviews Neuroscience*, ed. 2, 717-726, [online 2009.05.05], available at: <http://explore.georgetown.edu/people/michael/?action=viewpublications&PageTemplateID=129>

Underwood, G., Schmitt, N. and Galpin, A., 2004, "The eyes have it: An eye-movement study into the processing of formulaic sequences" in Schmitt, N. (ed.), 2004, *Formulaic Sequences: Acquisition, Processing and Use,* Amsterdam: John Benjamins, 153-172.

Vogel Sosa A. and MacFarlane J., 2002 "Evidence for frequency-based constituents in the mental lexicon: Collocations involving the word of", *Brain and Language,* vol. 83, 227–236.

Wray, A., 2000, “Formulaic sequences in second language teaching: principle and practice”, *Applied Linguistics,* vol. 21(4), 463-489.

Wray, A., 2002, *Formulaic Language and the Lexicon,* Cambridge: Cambridge University Press.

Wray, A., 2006, “Formulaic Language” in Brown, K. (ed.), *Encyclopedia of Language and Linguistics, 2nd edition,* Oxford: Elsevier, vol. 4, 590-597.

Wray, A. and Perkins, M., 2000, “The functions of formulaic language: an integrated model”, *Language & Communication,* vol. 20(1): 1-28.

Appendix I

Test sequences

Sequence nb./Dummy	Test sequences	D(ummy)/F(requent)/ I(n)F(requent)
A	det er ikke så greit	D
B	alle liker vel kake	D
C	kan jeg få være med	D
D	ja det er at det	D
E	det er jo bare å	D
F	de enkelte kapitlene	D
1	ja men det er jo	F
D	å skrive er en viktig	D
30	det er sånn jeg syns	IF
D	du lærer å skrive	D
2	for å si det sånn	F
D	gjennom tilbakemelding	D
31	og det var hvis du	IF
D	og samarbeid med andre	D
3	det er det som er	F
D	viktig å bli bevisst på	D
32	men det var jeg bare	IF
D	deg selv som skriver	D
4	nei jeg vet ikke jeg	F
D	om å velge ut hva	D
33	det er det å da	IF
D	du skal lese og orientere deg	D
*	holdt jeg på å si	F
D	om i en tekstkilde	D
34	at det er at jeg	IF
D	målrettet lesing	D
5	det er i hvert fall	F
D	doble notater	D
35	på en måte er sånn	IF
D	det kritiske sammendraget	D
6	jeg tror ikke det er	F
D	ulikt formål med teksten	D
36	det er så jeg ikke	IF
D	fra emne til problemstilling	D
7	et eller annet sånt noe	F
D	to strategier for	D
37	det var en jeg skal	IF
D	å utvikle strukturen i en tekst	D
8	det er ikke så veldig	F

D	å utdype moment i en	D
38	nei det var på det	IF
D	tekstutforskende skriving	D
9	jeg har lyst til å	F
D	noen karakteristiske trekk	D
39	er ikke noe på det	IF
D	uformell skriving spesielt	D
10	Og så er det jo	F
D	i feltarbeid og praksis	D
40	jeg tror det jeg var	IF
D	porteføljevurdering i kombinasjon	D
11	det er jo ikke det	F
D	med tradisjonelle prøveformer	D
41	nå er det som er	IF
D	grunnstrukturen formulert	D
12	i og med at jeg	F
D	fortellingen og den personlige	D
42	er jo det hvis du	IF
D	metaperspektiv på veiledningen	D
13	det er klart det er	F
D	avsnitt og logiske tekstmarkører	D
43	det har jeg som er	IF
D	den håndverksmessige siden	D
14	jeg er ikke helt sikker	F
D	vanlige formuleringer og	D
44	så var det jeg har	IF
D	begreper i oppgaveformuleringer	D
15	det er på en måte	F
D	oppgaveveiledning: store variasjoner	D
45	at det var hvis du	IF
D	samtaleregler og arbeidsgang ved	D
16	og da var det jo	F
D	tilbakemelding på skriftlige utkast	D
46	så har jeg en gang	IF
D	strategier for tilbakemelding	D
17	det er ikke noe problem	F
D	i ei skrivegruppe	D
47	ja og så er ikke	IF
D	tre fokus for responsen	D
18	nei jeg tror ikke det	F
D	den snakkende skriveren	D
48	for det er jeg bare	IF
D	responsgrupper via e-post eller	D
19	jeg er veldig glad i	F
D	elektroniske kommunikasjonsverktøy	D
49	er jo ikke på en	IF
D	råd til dem som vil	D

20	Holdt jeg på å si	F
D	vil starte skrivegruppe	D
50	og så er så mye	IF
D	Hjelpeliste for skrivere	D
21	det er det det er	F
D	samarbeidsskriving med IKT	D
51	jeg har ikke det at	IF
D	bevissthet om vurderingskriterier	D
22	ikke så veldig mye	F
D	videreutvikling av nybegynner	D
52	det er liksom at jeg	IF
D	innledning: figurer og troper	D
23	ja det er jo det	F
D	identisk gjentagelse	D
53	er jo det når jeg	IF
D	variert helgjentagelse	D
24	det er klart det	F
D	semantiske figurer	D
54	i hvert fall er sånn	IF
D	Noen ord til avslutning	D
25	og det syns jeg er	F
D	i en retorisk utsagnstype	D
55	er det ikke ja og	IF
D	det identisk gjentatte element	D
26	det er ikke noe sånn	F
D	de foregående tre hovedtyper	D
56	vet ikke jeg at det	IF
D	som ved de forskjellige	D
27	ja det er sant	F
D	rorlenken klirret og klapret	D
57	er det jo til å	IF
D	et fragment av noe stort	D
28	men det er jo det	F
D	i disse eksemplene fremtrer	D
58	da var det jeg hadde	IF
D	mine idealer har visnet	D
29	jeg syns det er veldig	F
D	to eller flere eller alle ledd	D
59	det er jo og det	IF

Appendix II

Infrequent target sequences in their original or constructed context.

Det er sånn jeg syns vi kristne burde møte jenter som er blitt uplanlagt gravid i stede for…
Jeg fikk et godt råd av en lege en gang **og det var: Hvis du** virkelig ikke er sulten, mistet totalt matlysten, så spis noe du virkelig har lyst på! **...**
Jeg hadde forøvrig ikke noen post partum-samtale heller, **men det var jeg bare** glad for, for da det hele endelig var over og jordmoren gikk, **...**
Det er det å da skulle sitte der i time etter time…
Lasse mener nok **at det er at jeg** sjelden gidder å lese korrektur på det jeg legger ut på nettet.
Det syns jeg **på en måte er sånn**... jeg vet ikke... jeg.
Det er så jeg ikke riktig VET hva jeg skal gjøre med all denne makten?
Han vi møtte: **det var en jeg skal** delta på kurs sammen med.
Nei, det var på det private budsjettet til en i vårt reportasjeteam.
Det **er ikke noe på det** nåværende tidspunkt som tyder på at det er noe mellom de to sakene, sier Utgaard
men **jeg tror det jeg var** redd var denne flokkmentaliteten.
Så **nå er det som er** godt for Wikipedia, godt for verden?
men den **er jo det hvis du** tar ut turtallssperren, og da kan ikke folk si at du direkte har trimmet bilen**...**
Du har ikke bart, men **det har jeg som er** trønder!
Hvis jeg tenker etter, **så var det jeg har** gjort ikke noen stor feil…
Jeg fortalte det **at det var hvis du** skulle være med på hytta, så måtte vi kjøre to biler…
Og **så har jeg en gang** tenkt å gi en drittsekk skikkelig som h*n fortjener…
Ja, og så er ikke det noe å skryte av, liksom?
For det er jeg bare så utrolig lei av…
Vi **er jo ikke på en** øde øy
Og så er så mye av det som sies bare tull.
Jeg har ikke det at det kiler i magen når jeg ser ham, og det savner jeg...
Det verste med det, **det er liksom at jeg** tenker mer på min egen rolle i meg selv, enn jeg

gjør på bøkene mine.
Det **er jo det når jeg** tenker meg om.
men selv jeg – som **i hvert fall er sånn** middels interessert i rock – foretrekker originalen, **...**
…så **er det ikke ja og** nei, men et entydig JA.
Og hvorfor **vet ikke jeg at det** er julaften, før nå?
men så lenge eg ikkje blir verre så **er det jo til å** leve med. **...**
Da var det jeg hadde så ”steikandes” lyst til å si:...
Tja, **det er jo òg det** der med at hun kanskje burde ha tatt litt ansvar selv.

Appendix III

All incorrect responses (responses occurring more then once is marked with gray):

Sq. nr.	Target sequence		Responce	Occur-rences	Cate-gory	English word-for-word translation	English equivalents
1.	Ja men det er jo	→	Jeg må lære å	1	5	I must learn to	'I must learn to'
"	"	→	Ja men det er det jo	1	2	Yes but it is it yes	'Yes but it is'
"	"	→	Man lærer jo	1	5	You learn yes	'You learn'
2.	For å si det sånn	→	Men hvordan	1	5	But how	'But how'
"	"	→	Borte i det danske	1	5	Away in the Danish	'Away in the Danish'
"	"	→	For å drive dank	2	4	For to drift loaf	'to idle about'
3.	Det er det som er	→	Jeg tror ikke det	1	5	I believe not it	'I don't think it is'
"	"	→	Jeg vet at det er	1	4	I know that it is	'I know that it is'
"	"	→	Det er jo det som er	1	2	It is yes it which is	'It is after all'
4.	Nei jeg vet ikke jeg	→	Jeg hører ikke	1	4	I hear not	'I can't hear'
"	"	→	Det er det jo ikke	1	5	It is it yes not	'Well it isn't'
"	"	→	Nei jeg vet ikke	8	3	No I know not	'No I don't know'
"	"	→	Det er jo ikke	1	5	It is yes not	'After all it's not'
"	"	→	Det er ett med lokket	1	5	It is one with lid	'It is one with the lid'
"	"	→	Nei det går ikke	1	4	No that goes not	'No hat won't go'
"	"	→	Jeg vet ikke jeg	1	3	I know not I	'I don't know'
"	"	→	Nei men ikke	1	4	No but not	'No but not'
"	"	→	Det gjør jeg ikke	1	4	It will I not	'I won't do it'
"	"	→	Det er heller ikke	1	5	It is either not	'It is nor'
5.	Det er i hvert fall	→	Jeg vil hvert fall	1	4	I will at least	'I will at least'
"	"	→	Det er det i hvert fall	2	2	It is it at least	'It is at least'
"	"	→	Jeg er i hvert fall	2	4	I am at least	'I am at least'
"	"	→	Diverss... sånn	1	5	?... so	'?... so'
6.	Jeg tror ikke det er	→	Dette er	1	5	This is	'This is'
"	"	→	Jeg tror ikke det er det	4	2	I believe not it is it	'I don't think it is'
"	"	→	Det er	2	3	It is	'It's'
"	"	→	Jeg tror ikke det er noe	1	2	I believe not it is something	'I don't think it is something'
"	"	→	Jeg tror det er	1	3	I believe it is	'I think it is'
"	"	→	Jeg tror at det er	1	4	I believe that it is	'I think it is'
7.	Et eller annet sånt noe	→	Fjernlale	1	5	Distant speech	'Distant speech'
"	"	→	Et eller annet	1	3	One or another	'Something'
"	"	→	Et eller annet så... sånn	1	4	One or another then ... like	'Something ... like'
"	"	→	Reklame	1	5	Advertising	'Advertising'
8.	Det er ikke så veldig	→	Jeg er ikke så veldig	2	4	I am not so very	'I am not very'

9.	Jeg har lyst til å	→	Jeg liker det	1	5	I like it	'I like it'
"	"	→	Jeg liker deg	1	5	I like you	'I like you'
"	"	→	Jeg har vel	1	4	I have vell	'I guess I have'
10.	Og så er det jo	→	Så er det jo	1	3	Then is it yes	'It is like that'
"	"	→	Nå er det jo	1	4	Now is it yes	'Then it is'
"	"	→	Nå er det noe	1	4	Now is it something	'It is something'
"	"	→	Så er det jo	1	3	So is it yes	'Then it is'
11.	Det er jo ikke det	→	Det er ikke der	1	4	It is not there	'It's not there'
"	"	→	Er jo ikke det	1	3	Is yes not it	'It's not'
12.	I og med at jeg	→	I media	2	5	In the media	'In the media'
"	"	→	Mediet	1	5	Media	'Media'
"	"	→	Går ned her	1	5	Goes down here	'Goes down here'
"	"	→	I og med	1	3	In and with	'It follows'
"	"	→	Nå vet jeg	1	5	Now know I	'Now I know'
"	"	→	Med det	1	5	With it	'Goes with'
"	"	→	Ho	1	5	Her	'Her'
13.	Det er klart det er	→	Det er klart at det er	2	2	It is clear that it is	'It is obvious that it is'
"	"	→	Det er viktig at det er	1	4	It is important that it is	'It is important that it is'
"	"	→	Du er klar over at det er	1	4	You are aware of that it is	'Are you aware of that it is'
"	"	→	Jeg er glad det er	1	4	I am happy it is	'I am happy it is'
"	"	→	Hvor interessant det er	1	4	How interesting it is	'How interesting it is'
"	"	→	Jeg er lei for at det er	1	4	I am sorry for that it is	'I am sorry that it is'
"	"	→	Eksamen er	1	5	The exam is	'The exam is'
"	"	→	Jeg er glad at det er	1	4	I am happy that it is	'I am happy that it is'
"	"	→	Det er bra at det er	1	4	It is good that it is	'It is nice that it is'
"	"	→	Jeg er glad det er	1	4	I am happy it is	'I am happy it is'
15.	Det er på en måte	→	Måte	2	3	Way	'How, manner'
"	"	→	Det med en måned	1	4	It with one month	'Regarding the one month'
"	"	→	Jeg er på en måte	1	4	I am on a way	'I am in some way'
"	"	→	Gjerne mange	1	5	Readily many	'Willingly many'
"	"	→	Gjerne	1	5	Certainly	'Certainly'
"	"	→	Gjerne med måte	1	5	Certainly with way	'I would like some moderation'
"	"	→	På en måte	3	3	In a way	'Somehow'
"	"	→	Det er med måte	1	4	It is with moderation	'It is with moderation'
16.	Og da var det jo	→	Og hva var det nå	1	4	And what was it now	'What was it about'
"	"	→	Det var det noe	1	4	It was it something	'It was something'

”	”	→	Og da var det noe	2	4	And then was it something	‘And then it was something’
”	”	→	Barn	1	5	Children	‘Children’
17.	Det er ikke noe problem	→	Seerne ser	1	5	The viewers see	‘The viewers see’
”	”	→	kjede	1	5	Chain	‘Chain’
”	”	→	Jeg er ikke noe interessert	1	4	I am not something interested	‘I am not slightly interested’
”	”	→	Det er ikke noe pent	1	4	It is not something beautiful	‘It is not beautiful at all’
”	”	→	Det er	1	3	It is	‘It is’
19.	Jeg er veldig glad i	→	Jeg er veldig glad i å	3	2	I am very happy in to	‘I am very fond of doing’
”	”	→	Jeg er veldig glad i deg	3	2	I am very fond in you	‘I am very fond of you’
”	”	→	Jeg er veldig glad i det	1	2	I am very glad in that	‘I like it very much’
”	”	→	Jeg er veldig glad i mat	1	2	I am very fond in food	‘I like food very much’
”	”	→	Jeg er veldig	1	3	I am very	‘I am very’
20.	Holdt jeg på å si	→	Hotellet blir	1	5	The hotel becomes	‘The hotel will be’
”	”	→	Forskjellen på de	1	5	The difference on them	‘The difference between them’
”	”	→	Håper jeg på å bli	1	4	Hope I on to be	‘I hope to become’
”	”	→	Hotell og bil	1	5	Hotel and car	‘Hotel and car’
”	”	→	Fra samme by	1	5	From same town	‘From the same town’
22.	Ikke så veldig mye	→	Ikke så lenge du vil det	1	4	Not so long you want it	‘Not as long as you want it’
”	”	→	Ikke så nøye på	1	4	Not so careful on	‘Not so careful’
”	”	→	Ikke å tenke så mye	1	4	Not to think so much	‘Not to think so much’
”	”	→	Ikke så lenge igjen	2	4	Not so long again	‘Not much time left’
”	”	→	Ikke så lenge	4	4	Not so long	‘Not as long’
”	”	→	Ikke så lenge… så mye	1	4	Not so long … so much	‘Not as long … so much’
”	”	→	Ikke så lenge siden	2	4	Not so long since	‘Not so long time ago’
”	”	→	Ikke noe å tenke på det	1	5	Not something to think on that	‘Not something to think about’
”	”	→	Ikke så	2	3	Not so	‘Not so’
”	”	→	Ikke så lenge så	1	4	Not so long so	‘Not so long as’
25.	Og det syns jeg er	→	Og det feltet er	1	4	And that department is	‘And that department is’
”	”	→	Dette er	2	5	This is	‘This is’
”	”	→	Og det syns jeg	1	3	And it believe I	‘And I think it is’
”	”	→	Og det første er	1	4	And the first is	‘And the first is’
”	”	→	Og det sjette er	1	4	And the sixth is	‘And the sixth is’
”	”	→	Og det… er	1	3	And it … is	‘And it … is’
26.	Det er ikke noe sånn	→	Det er ikke noe tvang	1	4	It is not any force	‘It is not compulsory ‘
”	”	→	Det er ikke noe sak	1	4	It is not any case	‘It is no matter’
”	”	→	Det er ikke noe sannhet	1	4	It is not any truth	‘That is no truth’
”	”	→	Det er ikke noe sant	1	4	It is not any true	‘It is not true’

”	”	→	Det er ikke noe	1	3	It is not something	‘It is nothing’
28.	Men det er jo det	→	Det er jo det	1	3	It is yes it	‘Of course, it is’
29.	Jeg syns det er veldig	→	Det syns jeg er veldig	6	4	It think I am very	‘I think that is very’
”	”	→	Det syns jeg virkelig	1	4	It think I really	‘I really think that’
”	”	→	Jeg syns så veldig	1	4	I think so very	‘I really think that’
”	”	→	begrep	1	5	Notion	‘Notion’
”	”	→	? der syns jeg veldig	1	4	? there think I very	’? there I really think’
”	”	→	Det… veldig	1	3	It … very	‘I … very’
”	”	→	Veldig	2	3	Very	‘Very’
”	”	→	Tyttebær veldig	1	5	cowberry very	‘cowberry very’
”	”	→	Jeg syns det veldig	1	3	I think it very	‘I think that very’
30.	Det er sånn jeg syns	→	Det er sånn jeg ser	2	4	It is so I see	‘That is the way I see it’
”	”	→	Det kan jeg ikke	5	4	It can I not	‘I cannot do that’
”	”	→	Det er sånn	1	3	It is so	‘It is like that’
”	”	→	Det er sånn jeg skrev det	1	4	It is so I wrote it	‘It is the way I wrote it’
”	”	→	Brikke kan jeg si	1	5	piece can I say	‘A piece I might say’
”	”	→	Sånn jeg skriver	1	4	way I write	‘The way I write’
”	”	→	Det er sånn jeg	2	3	It is so I	‘It is the way I am’
”	”	→	Mer har jeg ikke	1	5	More have I not	‘I don’t have more’
”	”	→	Det er sånn jeg har skrevet	1	4	It is so I have written	‘It is how I have written’
”	”	→	Det	1	3	It	‘It’
”	”	→	Det er sånn jeg ser det	1	4	It is so I see it	‘It is the way I see it’
”	”	→	Det er sånn jeg sier	1	4	It is so I say	‘It is how I say it’
”	”	→	Det er sånn jeg ser på	1	4	It is so I see on	‘It is the way I see it’
31.	Og det var hvis du	→	Og det har du	2	4	And it have you	‘And you got it’
”	”	→	Og det var	4	3	And it was	‘And it was’
”	”	→	Og det er viktig	1	4	And it is important	‘And it is important’
”	”	→	Det var	1	3	It was	‘It was’
”	”	→	Og det hørte du	1	4	And it heard you	‘And you heard it’
”	”	→	Og det vet du også	1	4	And it know you too	‘And you know it too’
”	”	→	Og det var dit du	1	4	And it was there you	‘And it was where you’
”	”	→	Og det	1	3	And it	‘And it’
”	”	→	Og det har du fått til	1	4	And it have you managed to	‘And you managed to do that’
”	”	→	Og det var dumt du	2	4	And it was stupid you	‘And that was stupid of you’
”	”	→	Og det har du rett i	1	4	And it have you right in	‘And you are right about that’
”	”	→	Og det er	1	4	And it is	‘And it is’
”	”	→	Og det hadde ikke du	1	4	And it had not you	‘And you wouldn’t have’

”	”	→	Og det var viktig	1	4	And it was important	‘And it was important’
32.	Men det var jeg bare	→	Men det var jo en	1	4	But it was yes one	‘But it was one though’
”	”	→	Men det var det bare	2	4	But it was it just	‘But it was just it’
”	”	→	Men det var jo bare	9	4	But it was yes just	‘But it was just that’
”	”	→	Men det har jeg bare	1	4	But it have I just	‘But I have just done/got it’
”	”	→	Men det må jeg bare	2	4	But it must I just	‘But I just have to’
”	”	→	Men	1	3	But	‘But’
”	”	→	Det var jeg jo	1	4	It was I yes	‘But it was me’
33.	Det er det å da	→	Det er det da	1	3	It is it then	‘That is the thing’
”	”	→	Det blir jo da	4	4	It will yes then	‘It will be then’
”	”	→	Det… det da	1	3	It … it then	‘It … it then’
”	”	→	Jeg ble jo da	3	5	I was yes then	‘I was then’
”	”	→	Jukse det blir jo da	1	4	To cheat it will yes then	‘To cheat will then be’
”	”	→	Det er da	1	3	It is then	‘It is then’
”	”	→	Jeg vil jo da	2	5	I want yes then	‘I would like to’
”	”	→	Det var det hun sa	1	4	It was it she said	‘It is was that she said’
”	”	→	Jeg blir jo glad	1	5	I become yes happy	‘I will of course be happy’
”	”	→	Det er jo det å da	1	2	It is yes it to then	‘well it is then to’
34.	At det er at jeg	→	Og det er at jeg	17	4	And it is that I	‘And it is that I’
”	”	→	Han ser at jeg	1	4	He sees that I	‘He sees that I’
”	”	→	Ofte opplever jeg at jeg	1	4	Often experience I that I	‘I often experience that I’
”	”	→	Og ser at jeg	1	4	And see that I	‘And I see that I’
”	”	→	Som ser at jeg	1	4	As see that I	‘Who sees that I’
”	”	→	Jeg tror at jeg	1	4	I believe that I	‘I think that I’
”	”	→	Og flere vet jeg	1	5	And more know I	‘And I know several who’
”	”	→	At jeg	1	3	That I	‘That I’
35.	På en måte er sånn	→	På en måte	2	3	On one way	‘In some way’
”	”	→	På en måte er det sånn	5	2	On one way is it so	‘It is some way like that’
”	”	→	På en måte en sånn	2	4	On one way one so	‘In some way one like that’
”	”	→	På en måte som	1	4	On one way as	‘In some way as’
36.	Det er så jeg ikke	→	Det der ser jeg ikke	1	4	It there see I not	‘It don’t see that’
”	”	→	Det er så	3	3	It is so	‘It might be’
”	”	→	Det er så jeg	1	3	It is so I	‘It is how I’
”	”	→	Det er som jeg ikke	1	4	It is as I not	‘It is like I havn’t’
”	”	→	De er så enkle	1	4	They are so simple	‘They are so simple’
37.	Det var en jeg skal	→	Det var en kar	1	4	It was a guy	‘It was a guy’
”	”	→	Det var en vikar	1	4	It was a substitute	‘It was a substitute’

”	”	→	Det var	2	3	It was	‘It was’
”	”	→	Det var en kald	1	4	It was a cold	‘It was a cold’
”	”	→	Det var enighet fra oss	1	4	It was agreement from us	‘We all agreed’
”	”	→	Det var helt sentralt	1	4	It was totally vital	‘It was completely vital’
”	”	→	En hva jeg skal	1	4	One what I will	‘Someone I will’
”	”	→	Det var enighet om	1	4	It was agreement about	‘It was agreement about’
”	”	→	Det var jeg som skal	1	4	It was I who will	‘It will be me who will’
”	”	→	Det var en ifra	1	4	It was someone from	‘It was someone from’
”	”	→	Det var en sak	1	4	It was a case	‘It was a case’
”	”	→	Det var jeg som	1	4	It was I who	‘It was I who’
”	”	→	Det var en gang	1	4	It was one time	‘Once upon a time’
”	”	→	Det var en	1	3	It was one	‘It was one’
38.	Nei det var på det	→	Er det vannet	3	5	Is it the water	‘Is it the water’
”	”	→	Nei det	1	3	No it	‘No that’
”	”	→	Er det	3	5	Is it	‘Is it’
”	”	→	Nei det var ikke	1	4	No it was not	‘Not it was not’
”	”	→	Var på det	2	3	Was on that	‘Went to it’
”	”	→	Nei det var det ikke	2	4	No it was it not	‘No, it wasn’t it’
”	”	→	Er det bable	1	5	Is it babble	‘Is it nonsense’
”	”	→	Jeg er vant til det	1	5	I am used to it	‘I am used to it’
”	”	→	Er det vante	1	5	Is it the regular	‘It is the regular’
”	”	→	Hva var det for noe igjen	1	5	What was it for something again	‘What was that about’
”	”	→	Nei det var ikke det	2	4	No it was not it	‘No, it wasn’t’
39.	Er ikke noe på det	→	Er ikke noe for det	2	4	Is not something for it	‘It is nothing of it’
”	”	→	Det er ikke noe med det	5	4	It is not something with it	‘It has nothing to do with it’
”	”	→	Det er ikke noe det	1	4	It is not something it	‘It is nothing’
”	”	→	Det er ikke noe på det	6	2	It is not something on it	‘It is nothing’
”	”	→	Er ikke noe med det	4	4	Is not something with it	‘It has nothing to do with it’
”	”	→	Det er ikke noe av det	1	4	It is not something of it	‘It is nothing of it’
”	”	→	Er ikke noe problem	1	4	Is not one problem	‘It is not a problem’
”	”	→	Det er ikke noe med det	1	4	It is not something with it	‘It is nothing to do with that’
40.	Jeg tror det jeg var	→	Jeg tror det var	7	3	I believe it was	‘I believe it was’
”	”	→	Og det var	1	4	And it was	‘And it was’
”	”	→	Jeg trodde jeg var	1	4	I thought I was	‘I thought I was’
”	”	→	Det tror jeg det var	1	4	It believe I it was	‘I believe it was’
”	”	→	Det var	3	3	It was	‘It was’

”	”	→	Jeg tror ikke jeg har	1	4	I believe not I have	‘I don’t believe I have’
”	”	→	Nettopp det jeg var	2	4	Exactly it I was	‘I was exactly that’
”	”	→	Nettopp det tror jeg var	1	4	Exactly it believe I was	‘Exactly that I think was’
”	”	→	Det skulle det det var	1	4	It should it it was	‘Exactly that it was ought to be’
”	”	→	Det jeg har	1	4	It I have	‘I have it’
”	”	→	Var	1	3	Was	‘Was’
”	”	→	Jeg syns det var	1	4	I think it was	‘I think it was’
41.	Nå er det som er	→	Det er det som er	14	4	It is it as is	‘It is exactly that which is’
”	”	→	Ta det som er	1	4	Take it as is	‘Take it as it is’
”	”	→	Har den som er… var	1	4	Has it as is … was	‘Does it have who is … was’
”	”	→	Hva er det som er	4	4	What is it as is	‘What is it which is like’
”	”	→	Tar det som er	1	4	Take it as is	‘Take what is left’
”	”	→	Jeg har vent med	1	5	I have used with	‘I have got used to it’
”	”	→	Det som er	1	3	It as is	‘What is’
42.	Er jo det hvis du	→	Er jo det vet du	4	4	Is yes it know you	‘You know it is’
”	”	→	Det er jo det hvis du	4	2	It is yes it if you	‘It would be like that if you’
”	”	→	Er jo det som de	1	4	Is yes it as they	‘It is what they’
”	”	→	Det er jo det vet du	3	4	It is yes it know you	‘You know it is that’
43.	Det har jeg som er	→	Det har jeg for nær	3	4	It have I too close	‘That is too close to me’
”	”	→	Det har jeg	3	3	It have I	‘I have that’
”	”	→	Det har jeg for nære	2	4	It have I too close	‘I have it too close’
”	”	→	Jeg var så nær	1	5	I was so close	‘I was so close’
”	”	→	Det har jeg så nære	1	4	It have I so close	‘I have that so close’
”	”	→	Det tar jeg for nær	1	4	It take I too close	‘I take that too close’
”	”	→	Det har jeg som nær	1	4	It have I too close	‘I got that close to me’
”	”	→	Da var jeg for nære	2	4	When was I too close	‘When I was too close’
”	”	→	Det har jeg ikke her	1	4	It have I not here	‘I don’t have it here’
”	”	→	Jeg har deg for nære	1	5	I have you too close	‘You are just too close to me’
”	”	→	Det har jeg på nær	1	4	It have I on close	‘I have that too close’
”	”	→	Det tar jeg meg nær av	1	4	It take I me close of	‘It hurts me’
”	”	→	Det var jeg som	1	4	It was I who	‘It was me who’
44.	Så var det jeg har	→	Så det er det jeg har	1	4	So it is it I have	‘So this is what I got’
”	”	→	Som er de jeg har	1	4	As is they I have	‘As they I got’
”	”	→	Så var det jeg var	4	4	So was it I was	‘I was like that’
”	”	→	Det var det jeg har	4	4	It was it I have	‘This is what I got’
”	”	→	Det var det jeg sa	1	4	It was it I said	‘That was what I said’

”	”	→	Så det er det jeg har	3	4	So it is it I have	‘So this is what I got’
”	”	→	Så var det det jeg har	1	2	So was it it I have	‘So this was what I got’
”	”	→	Som er det jeg har	1	4	As is it I have	‘Which is what I got’
”	”	→	Det er bare det jeg har	1	4	It is just it I have	‘It is just what I got’
”	”	→	Det var det jeg var	1	4	It was it I was	‘It was what I used to be’
”	”	→	Så det var det jeg har	1	2	So it was it I have	‘So it was what I used to have’
”	”	→	Så var det	1	3	So was it	‘It was like that’
”	”	→	Så er det jeg har	1	4	So is it I have	‘So this is what I got’
45.	At det var hvis du	→	Og det er vanskelig	1	5	And it is difficult	‘And it is difficult’
”	”	→	Ja det var visst lurt	2	4	Yes it was certainly smart	‘That was certainly clever’
”	”	→	Det var visst lurt	2	4	It was certainly smart	‘It was certainly clever’
”	”	→	At det var ikke så lurt	1	4	That it was not so smart	‘That it was not that clever’
”	”	→	At det var du visst tur	1	4	That it was you certainly trip	‘That is was certainly you trip’
”	”	→	Kan se vanskelig ut	1	5	Can look difficult out	‘It may look difficult’
”	”	→	Det var det	1	4	It was it	‘It was it’
”	”	→	Ja det var hvis du	1	4	Yes it was if you	‘Yes, in case you would’
”	”	→	Og det var hvis du	2	4	And it was if you	‘And that would be if you’
”	”	→	At det var	1	3	That it was	‘That it was’
”	”	→	At det var vel surt	1	4	That it was too sour	‘That it was too sour’
”	”	→	Ja det var visst	1	4	Yes it was certainly	‘Yes it was certainly’
”	”	→	Skal se da hvis du	1	4	Should see then if you	‘Will see then if you’
46.	Så har jeg en gang	→	Så har jeg egne	1	4	So have I my own	‘So I have my own’
”	”	→	Det var jeg en gang	2	4	It was I one time	‘That was me once’
”	”	→	Ofte har jeg en gang	1	4	Often have I one time	‘I often have’
”	”	→	Og så har jeg en gang	1	4	And so have I one time	‘And I have once’
”	”	→	Det har jeg en gang	2	4	It have I one time	‘I have done that once’
”	”	→	Da har jeg en gang	1	4	When have I one time	‘Once I did’
”	”	→	Så har jeg en klar	1	4	So have I one clear	‘I have one clear’
”	”	→	En gang	1	3	One time	‘Once’
”	”	→	Det tar jeg en gang	1	4	It takes I one time	‘I will take that sometime’
”	”	→	Det var jeg egentlig	1	5	It was I really	‘I really was that’
47.	Ja og så er ikke	→	Ja og så har vi det	1	4	Yes and so have we it	‘Yes, and we have it’
”	”	→	Ja så har man ikke	1	4	Yes so has one not	‘One doesn’t have it like that’
”	”	→	Ja sånn… så er det ikke	1	4	Yes like … so is it not	‘Like that... not like that’
”	”	→	Ja så	5	3	Yes so	‘Right then’
”	”	→	Ja så har vi ikke	3	4	Yes so have we not	‘We don’t have it like that’
”	”	→	Å ja og så hans	1	4	To yes and so his	‘Oh yes, and his’

”	”	→	Ja så var det	1	4	Yes so was it	‘It was like that’
”	”	→	Ja sa han til det	1	4	Yes said him to that	‘He agreed to do it’
”	”	→	Ja så har jeg ikke	1	4	Yes so have I not	‘I don’t have it like that’
”	”	→	Ja så… det	1	4	Yes so … it	‘Right … that’
”	”	→	Ja så har vi ikke	2	4	Yes so have we not	‘We don’t have it like that’
”	”	→	Ja sa hauken	1	5	Yes said the hawk	‘The hawk said yes’
”	”	→	Ja så er det ikke	1	4	Yes so is it not	‘Well, it is not like that’
”	”	→	Ja se	1	5	Yes see	‘Yes, look’
”	”	→	Ja og så har vi ikke	1	4	Yes and so have we not	‘And we don’t have it like that’
”	”	→	Ja og så er det ikke	1	4	Yes and so is it not	‘And it is not like that’
”	”	→	Ja så har vi	1	4	Yes so have we	‘This is how we’
48.	For det er jeg bare	→	For det er å være	1	4	For it is to be	‘It is to be’
”	”	→	For det er jo bare	17	4	For it is yes just	‘It is just to’
”	”	→	Jeg bare	1	3	I just	‘I would just’
”	”	→	Tror jo bare	1	5	Believe yes just	‘Only thinks’
”	”	→	For det er bare	1	4	For it is just	‘It is just’
”	”	→	Det er jo bare	1	4	It is yes just	‘It is just’
”	”	→	For det er bare	1	4	For it is just	‘It is just’
49.	Er jo ikke på en	→	Er du på en	1	4	Are you on one	‘Are you on it’
”	”	→	Jeg lurer på om	1	5	I wonder on if	‘I wonder whether’
”	”	→	Det er jo ikke noe	1	4	It is yes not one	‘It is nothing’
”	”	→	Er vi ikke på en	1	4	Are we not on one	‘Are we not on it’
”	”	→	Er du ikke på en	1	4	Are you not on one	‘Are you not on it’
”	”	→	Er jo ikke noe	2	4	Is yes not one	‘It is nothing’
”	”	→	Er gjort på en	1	4	Is done on one	‘It is done in one’
”	”	→	Er jo ikke	1	3	Is yes not	‘Is not’
”	”	→	På en	2	3	On one	‘On a’
50.	Og så er så mye	→	Så det er så mye	1	4	So it is so much	‘So it is so much’
”	”	→	Hvis det er så mye	3	4	If it is so much	‘If it is so much’
”	”	→	Visste ikke at det var så mye	1	4	Knew not that it was so much	‘Didn’t know it was so much’
”	”	→	Det er så mye	1	4	It is so much	‘It is so much’
”	”	→	mye	1	3	Much	‘Much’
”	”	→	Hvis du har så mye	2	4	If you have so much	‘If you have so much’
”	”	→	Det spørs hvor mye	1	5	It depends how much	‘It depends on how much’
”	”	→	Du er så mye	1	4	You are so much	‘You are so full of’
”	”	→	Det var så mye	2	4	It was so much	‘It was so much’

”	”	→	Som er så mye	1	4	As is so much	‘Which is so much’
”	”	→	Så er så mye	2	3	So is so much	‘This is so much’
”	”	→	Og så er det så mye	1	2	And so is it so much	‘And then it is so much’
”	”	→	Så mye	5	3	So much	‘So much’
”	”	→	Koster bare så mye	1	4	Costs just so much	‘It only costs this much’
”	”	→	Det som var så mye	1	4	It as was so much	‘Which was so much’
”	”	→	Noe var så mye	1	4	Something was so much	‘Something was so much’
51.	Jeg har ikke det at	→	Jeg har ikke det her	1	4	I have not it here	‘I don’t have it here’
”	”	→	Jeg har ikke det	9	3	I have not it	‘I don’t have it’
”	”	→	Jeg hadde ikke det ja	2	4	I have not it yes	‘I didn’t have it’
”	”	→	Jeg hadde ikke det da	1	4	I had not it then	‘I didn’t have it then’
”	”	→	Jeg har ikke det da	1	4	I have not it then	‘I didn’t have it then’
”	”	→	Jeg har ikke det og	1	4	I have not it and	‘I don’t have that either’
”	”	→	Nå er det ikke det	1	4	Now is it not it	‘It is not that’
”	”	→	Jeg har ikke det jeg	2	4	I have not it I	‘I don’t have it’
”	”	→	Jeg har jo ikke det	1	4	I have yes not it	‘I don’t have it’
52.	Det er liksom at jeg	→	Det er litt sånn at jeg	2	4	It is little like that I	‘It is a bit like if I’
”	”	→	Det er liksom jeg	9	3	It is somewhat I	‘It is somewhat me’
”	”	→	Det er lenge at jeg	1	4	It is long that I	‘It is a long time that I’
”	”	→	Det ble liksom at jeg	1	4	It became somewhat that I	‘It went somehow that I’
”	”	→	Det er litt at jeg	5	4	It is a little that I	‘It is a bit like I’
”	”	→	Det er… jeg	1	3	It is … I	‘It is … I’
”	”	→	Det er litt satt jeg	1	4	It is little sat I	‘It is a little sat I’
53.	Er jo det når jeg	→	Det er jo det når jeg	15	2	It is yes it when I	‘It is like that when I’
”	”	→	Du kan jo det når jeg	1	4	You can yes it when I	‘You can do it when I’
”	”	→	Det er jo den jeg	1	4	It is yes that I	‘It is that I’
54.	I hvert fall er sånn	→	I alle fall er det sånn	1	4	In all case is it so	‘In any case is it like this’
”	”	→	Hvert fall er det sånn	2	3	any case is it so	‘In any case is it like this’
”	”	→	Hvert fall er sånn	3	3	any case is so	‘In any case is it like this’
”	”	→	Hvert fall det er sånn	1	4	any case it is so	‘In any case is it like this’
”	”	→	I hvert fall er det sånn	2	2	In any case is it so	‘In any case is it like this’
”	”	→	Han er så	1	5	He is so	‘He is so’
”	”	→	Er sånn	1	3	Is so	‘Is like that’
”	”	→	Lakken er slang	1	5	The varnish is slang	‘The varnish is slang’
”	”	→	Slå	1	5	Beat	‘Beat’
55.	Er det ikke ja og	→	Er det ikke ja og nei	1	2	Is it not yes and no	‘Isn’t it yes and no’
”	”	→	Er ikke ja og	1	3	Is not yes and	‘Isn’t yes and’

”	”	→	Takke ja og	1	4	Thank yes and	‘To say thank you and’
”	”	→	Og piano	1	5	And piano	‘And piano’
”	”	→	Er det ikke ja	1	3	Is it not yes	‘Isn’t it a yes’
”	”	→	Er Pia og	1	4	Is Pia and	‘Is Pia and’
”	”	→	Er ikke ja	1	3	Is not yes	‘It is not yes’
”	”	→	Ja og	2	3	Yes and	‘Yes and’
56.	Vet ikke jeg at det	→	Det fjerde at det	1	4	The fourth that it	‘The fourth is that’
”	”	→	Tror ikke jeg at det	2	4	Believe not I that it	‘I don’t believe that it’
”	”	→	Syns ikke jeg at det	1	4	Find not I that it	‘I don’t find it’
”	”	→	Tenker jeg at det	1	4	Think I that it	‘I think that it’
”	”	→	Det fjerde av	1	5	The fourth of	‘The fourth of’
”	”	→	Det fjerde er at det	1	4	The fourth is that it	‘The fourth is that it’
”	”	→	Det vet ikke jeg altså	1	4	It know not I right	‘I don’t know that right’
”	”	→	Vet ikke jeg at	1	3	Know not I that	‘I don’t know that’
”	”	→	Jeg at det	1	3	I that it	‘I that it’
”	”	→	Jeg at	1	3	I that	‘I that’
”	”	→	Visste ikke jeg at det	1	4	Knew not I that it	‘I didn’t know that I’
”	”	→	Det viktigste er at det	1	4	The most important is that it	‘The most important is that it’
”	”	→	Vet ikke jeg altså	1	4	Knew not I right	‘I don’t know that right’
”	”	→	Det tror jeg at det	1	4	It believe I that it	‘I believe that it will’
57.	Er det jo til å	→	Er du full av	2	5	Are you full of	‘Are you full of’
”	”	→	Da er det jo fint å	1	4	Then is it yes nice to	‘Then is it nice to’
”	”	→	Er du snill å	1	5	Are you kind to	‘Are you kind to’
”	”	→	Er du	1	5	Are you	‘Are you’
58.	Da var det jeg hadde	→	Det var det jeg hadde	11	4	It was it I had	‘It was all I had’
”	”	→	Hva var det jeg hadde	11	4	What was it I had	‘What was it I had’
59.	Det er jo òg det	→	Det er lov det	7	4	It is permitted that	‘That is permitted’
”	”	→	Det er nå det	6	4	It is now it	‘It is now that’
”	”	→	Det er noe i det	1	4	It is something in that	‘It is something about that’
”	”	→	Jeg må det	4	5	I must it	‘I must do it’
”	”	→	Det er jo nå det	2	4	It is yes now it	‘It is now that’
”	”	→	Det er noe det	1	4	It is something it	‘It is something about that’
”	”	→	Det er noe med det	1	4	It is something with that	‘It is something about that’
”	”	→	Det er noe av det	1	4	It is something of that	‘It is something about that’
”	”	→	Det er jo og noe det	1	2	It is yes and something it	‘It is something about that’
”	”	→	Det er da det	1	4	It is then it	‘It is something’